Proceedings of the WILDRE5– 5th Workshop on Indian Language Data: Resources and Evaluation, pages 7–13
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

A Dataset for Troll Classification of TamilMemes

**Shardul Suryawanshi[1], Bharathi Raja Chakravarthi[1], Pranav Varma[2] ,
Mihael Arcan[1], John P. McCrae[1] and Paul Buitelaar [1]**
[1] Insight SFI Research Centre for Data Analytics
[1]Data Science Institute, National University of Ireland Galway
[2]National University of Ireland Galway
{shardul.suryawanshi, bharathi.raja}@insight-centre.org

Abstract

Social media are interactive platforms that facilitate the creation or sharing of information, ideas or other forms of expression among people. This exchange is not free from offensive, trolling or malicious contents targeting users or communities. One way of trolling is by making memes, which in most cases combines an image with a concept or catchphrase. The challenge of dealing with memes is that they are region-specific and their meaning is often obscured in humour or sarcasm. To facilitate the computational modelling of trolling in the memes for Indian languages, we created a meme dataset for Tamil (TamilMemes). We annotated and released the dataset containing suspected trolls and not-troll memes. In this paper, we use the a image classification to address the difficulties involved in the classification of troll memes with the existing methods. We found that the identification of a troll meme with such an image classifier is not feasible which has been corroborated with precision, recall and F1-score.

Keywords: Tamil dataset, memes classification, trolling, Indian language data

1. Introduction

Traditional media content distribution channels such as television, radio or newspapers are monitored and scrutinized for their content. Nevertheless, social media platforms on the Internet opened the door for people to contribute, leave a comment on existing content without any moderation. Although most of the time, the internet users are harmless, some produce offensive content due to anonymity and freedom provided by social networks. Due to this freedom, people are becoming creative in their jokes by making memes. Although memes are meant to be humorous, sometimes it becomes threatening and offensive to specific people or community.

On the Internet, a troll is a person who upsets or starts a hatred towards people or community. Trolling is the activity of posting a message via social media that is intended to be offensive, provocative, or menacing to distract which often has a digressive or off-topic content with the intent of provoking the audience (Bishop, 2013; Bishop, 2014; Mojica de la Vega and Ng, 2018; Suryawanshi et al., 2020). Despite this growing body of research in natural language processing, identifying trolling in memes has yet to be investigated. One way to understand how meme varies from other image posts was studied by Wang and Wen (2015). According to the authors, memes combine two images or are a combination of an image and a witty, catchy or sarcastic text. In this work, we treat this task as an image classification problem.

Due to the large population in India, the issue has emerged in the context of recent events. There have been several threats towards people or communities from memes. This is a serious threat which shames people or spreads hatred towards people or a particular community (Kumar et al., 2018; Rani et al., 2020; Suryawanshi et al., 2020). There have been several studies on moderating trolling, however, for a social media administrator memes are hard to monitor as they are region-specific. Furthermore, their meaning is often obscure due to fused image-text representation. The content in Indian memes might be written in English, in a native language (native or foreign script), or in a mixture of languages and scripts (Ranjan et al., 2016; Chakravarthi et al., 2018; Jose et al., 2020; Priyadharshini et al., 2020; Chakravarthi et al., 2020a; Chakravarthi et al., 2020b). This adds another challenge to the meme classification problem.

(a) Example 1

(b) Example 2

Figure 1: Examples of Indian memes.

In Figure 1, Example 1 is written in Tamil with two images and Example 2 is written in English and Tamil (Roman Script) with two images. In the first example, the meme is trolling about the *"Vim dis-washer"* soap. The information in Example 1 can be translated into English as *"the price of a lemon is five Rupees"*, whereby the image below shows a crying person. Just after the crying person the text says *"The price of a Vim bar with the power of 100 Lemon is just 10 Rupees"*. This is an example of opinion manipulation with trolling as it influences the user opinion about products, companies and politics. This kind of memes might be effective in two ways. On the one hand, it is easy for companies and political parties to gain popularity. On the other hand, the trolls can damage the reputation of the company name or political party name. Example 2 shows a funny meme; it shows that a guy is talking to a poor lady while the girl in the car is looking at them. The image below includes a popular Tamil comedy actor with a short text written beneath *"We also talk nicely to ladies to get into a relationship"*.

Even though there is a widespread culture of memes on the Internet, the research on the classification of memes is not studied well. There are no systematic studies on classifying memes in a troll or not-troll category. In this work, we describe a dataset for classifying memes in such categories. To do this, we have collected a set of original memes from volunteers. We present baseline results using convolutional neural network (CNN) approaches for image classification. We report our findings in precision, recall and F-score and publish the code for this work at `https://github.com/sharduls007/TamilMemes`.

2. Troll Meme

A troll meme is an implicit image that intents to demean or offend an individual on the Internet. Based on the definition "Trolling is the activity of posting a message via social media that tend to be offensive, provocative, or menacing (Bishop, 2013; Bishop, 2014; Mojica de la Vega and Ng, 2018)". Their main function is to distract the audience with the intent of provoking them. We define troll memes as a meme, which contains offensive text and non-offensive images, offensive images with non-offensive text, sarcastically offensive text with non-offensive images, or sarcastic images with offensive text to provoke, distract, and has a digressive or off-topic content with intend to demean or offend particular people, group or race.

Figure 2 shows examples of trolling memes, Example 3 is trolling the potato chip brand called Lays. The translation of the text is *"If you buy one packet of air, then 5 chips free"*, with its intention to damage the company's reputation. Figure 2 illustrates examples of not-troll memes. The translation of Example 4 would be *"Sorry my friend (girl)"*. As this example does not contain any provoking or offensive content and is even funny, it should be listed in the not-troll category.

As a troll meme is directed towards someone, it is easy to find such content in the comments section or group chat of social media. For our work, we collected memes from volunteers who sent them through WhatsApp, a social media for chatting and creating a group chat. The suspected troll

(a) Example 3

(b) Example 4

Figure 2: Examples of troll and not-troll memes.

memes then have been verified and annotated manually by the annotators. As the users who sent these troll memes belong to the Tamil speaking population, all the troll memes are in Tamil. The general format of the meme is the image and Tamil text embedded within the image.

Most of the troll memes comes from the state of Tamil Nadu, in India. The Tamil language, which has 75 million speakers,[1] belongs to the Dravidian language family (Rao and Lalitha Devi, 2013; Chakravarthi et al., 2019a; Chakravarthi et al., 2019b; Chakravarthi et al., 2019c) and is one of the 22 scheduled languages of India (Dash et al., 2015). As these troll memes can have a negative psychological effect on an individual, a constraint has to be in place for such a conversation. In this work, we are attempting to identify such troll memes by providing a dataset and image classifier to identify these memes.

3. Related Work

Trolling in social media for text has been studied extensively (Bishop, 2013; Bishop, 2014; Mojica de la Vega and Ng, 2018; Malmasi and Zampieri, 2017; Kumar et al., 2018; Kumar, 2019). Opinion manipulation trolling (Mihaylov et al., 2015b; Mihaylov et al., 2015a), troll comments in News Community (Mihaylov and Nakov, 2016), and the role of political trolls (Atanasov et al., 2019) have been studied. All these considered the trolling on text-only media. However, meme consist of images or images with text.

[1] https://www.ethnologue.com/language/tam

A related research area is on offensive content detection. Various works in the recent years have investigated Offensive and Aggression content in text (Clarke and Grieve, 2017; Mathur et al., 2018; Nogueira dos Santos et al., 2018; Galery et al., 2018). For images, Gandhi et al. (2019) deals with offensive images and non-compliant logos. They have developed a computer-vision driven offensive and non-compliant image detection algorithm that identifies the offensive content in the image. They have categorized images as offensive if it has nudity, sexually explicit content, abusive text, objects used to promote violence or racially inappropriate content. The classifier takes advantage of a pre-trained object detector to identify the type of object in the image and then sends the image to the unit which specializes in detecting objects in the image. The majority of memes do not contain nudity or explicit sexual content due to the moderation of social media on nudity. Hence unlike their research, we are trying to identify troll memes by using image features derived by use of a convolutional neural network.

Hate speech is a subset of offensive language and datasets associated with hate speech have been collected from social media such as Twitter (Xiang et al., 2012), Instagram (Hosseinmardi et al., 2015), Yahoo (Nobata et al., 2016), YouTube (Dinakar et al., 2012). In all of these works, only text corpora have been used to detect trolling, offensive, aggression and hate speech. Nevertheless, for memes, there is no such dataset. For Indian language memes, it is not available as to our knowledge. We are the first to develop a meme dataset for Tamil, with troll or not-troll annotation.

4. Dataset

4.1. Ethics

For our study, people provided memes voluntarily for our research. Additionally, all personal identifiable information such as usernames are deleted from this dataset. The annotators were warned about the trolling content before viewing the meme, and our instructions informed them that they could quit the annotation campaign anytime if they felt uncomfortable.

4.2. Data collection

To retrieve high-quality meme data that would likely to include trolling, we asked the volunteers to provide us with memes that they get in their social media platforms, like WhatsApp, Facebook, Instagram, and Pinterest. The data was collected between November 1, 2019, until January 15, 2019, from sixteen volunteers. We are not disclosing any personal information of the volunteers such as gender as per their will. Figure 3 shows an example of the collected memes. We removed duplicate memes, however, we kept memes that uses the same image but different text. This was a challenging task since the same meme could have different file names. Hence the same meme could be annotated by different annotators. Due to this, we checked manually and removed such duplicates before sending them to annotators. An example is shown in Figure 3, where the same image with different text is used. Example 5 describes the image as *"can not understand what you are saying"*, whereby Example 6 describes image as *"I am confused"*.

(a) Example 5

(b) Example 6

Figure 3: Examples on same image with different text.

4.3. Annotation

After we obtained the memes, we presented this data to the annotators using Google Forms. To not over-burden the annotators, we provided ten memes per page and hundred memes per form. For each form, the annotators are asked to decide if a given meme is of category troll or not-troll. As a part of annotation guidelines, we gave multiple examples of troll memes and not-troll memes to the annotators. The annotation for these examples has been done by the an annotator who is considered as a expert as well as a native Tamil speaker. Each meme is assigned to two different annotators, a male and a female annotator. To ensure the quality of the annotations and due to the region-specific nature of the annotation task, only native speakers from Tamil Nadu, India were recruited as annotators. Although we are not disclosing the gender demographics of volunteers who provided memes, we have gender-balanced annotation since each meme has been annotated by a male and a female.A meme is considered as troll only when both of the annotators label it as a troll.

4.4. Inter-Annotator Agreement

In order to evaluate the reliability of the annotation and their robustness across experiments, we analyzed the inter-annotator agreement using Cohen's kappa (Cohen, 1960). It compares the probability of two annotators agreeing by chance with the observed agreement. It measures agreement expected by chance by modelling each annotator with separate distribution governing their likelihood of assigning

a particular category. Mathematically,

$$K = \frac{p(A) - p(E)}{1 - p(E)} \quad (1)$$

where K is the kappa value, $p(A)$ is the probability of the actual outcome and $p(E)$ is the probability of the expected outcome as predicted by chance (Bloodgood and Grothendieck, 2013). We got a kappa value of 0.62 between two annotators (gender balance male and female annotators). Based on Landis and Koch (1977) and given the inherent obscure nature of memes, we got fair agreement amongst the annotators.

4.5. Data Statistics

We collected 2,969 memes, of which most are images with text embedded on them. After the annotation, we learned that the majority (1,951) of these were annotated as troll memes, and 1,018 as not-troll memes. Furthermore, we observed that memes, which have more than one image have a high probability of being a troll, whereas those with only one image are likely to be not-troll. We included Flickr30K[2] images (Young et al., 2014) to the not-troll category to address the class imbalance. Flickr30K is only added to training, while the test set is randomly chosen from our dataset. In all our experiments the test set remains the same.

5. Methodology

To demonstrate how the given dataset can be used to classify troll memes, we defined two experiments with four variations of each. We measured the performance of the proposed baselines by using precision, recall and F1-score for each class, i.e. "troll and not-troll". We used ResNet (He et al., 2016) and MobileNet (Howard et al., 2017) as a baseline to perform the experiments. We give insights into their architecture and design choices in the sections below.

ResNet

ResNet has won the ImageNet ILSVRC 2015 (Russakovsky et al., 2015) classification task. It is still a popular method for classifying images and uses residual learning which connects low-level and high-level representation directly by skipping the connections in-between. This improves the performance of ResNet by diminishing the problem of vanishing gradient descent. It assumes that a deeper network should not produce higher training error than a shallow network. In this experiment, we used the ResNet architecture with 176 layers. As it was trained on the ImageNet task, we removed the classification (last) layer and used *GlobalAveragePooling* in place of fully connected layer to save the computational cost. Later, we added four fully connected layers with the classification layer which has a sigmoid activation function.This architecture is trained with or without pre-trained ImageNet weights.

MobileNet

We trained MobileNet with and without ImageNet weights. The model has a depth multiplier of 1.4, and an input dimension of 224×224 pixels. This provides a $1,280 \times 1.4 = 1,792$ -dimensional representation of an image, which is then passed through a single hidden layer of a dimensionality of $1,024$ with ReLU activation, before being passed to a hidden layer with input dimension of (512,None) without any activation to provide the final representation h_p. The main purpose of MobileNet is to optimize convolutional neural networks for mobile and embedded vision applications. It is less complex than ResNet in terms of number of hyperparameters and operations. It uses a different convolutional layer for each channel, this allows parallel computation on each channel which is Depthwise Separable Convolution. Later on the features extracted from these layers have been combined using the pointwise convolution layer. We used MobileNet to reduce the computational cost and compare it with the computationally intensive ResNet.

[2]https://github.com/BryanPlummer/flickr30k_entities

6. Experiments

We experimented with ResNet and MobileNet. The variation in experiments comes in terms of the data on which the models have been trained on, while the test set (300 memes) remained the same for all experiments. In the first variation, ***TamilMemes*** in Table 1, we trained the ResNet and MobileNet models on our Tamil meme dataset(2,669 memes). The second variation, i.e. ***TamilMemes + ImageNet*** uses pre-trained ImageNet weights on the Tamil memes dataset. To address the class imbalance, we added 1,000 images from the Flickr30k dataset to the training set in the third variation i.e. ***TamilMemes + ImageNet + Flickr1k***. As a result, the third variation has 3,969 images (1,951 trolls and 2,018 not-trolls). In the last variation, ***TamilMemes + ImageNet + Flickr30k***, we added 30,000 images from the Flickr30k dataset to not-troll category. Flickr dataset has images and the captions which describes the image. We used these images as a not-troll category because they do not convey trollings without the context of the text. Except for the *TamilMemes* baseline, we are using pre-trained ImageNet weights for all other variations. Images from the Flickr30k dataset are used to balance the not-troll class in the *TamilMemes + ImageNet + Flickr1k* variation. On the one hand, the use of all the samples from the Flickr30k dataset as not-troll in the fourth variation introduces the class imbalance by significantly increasing the number of not-troll samples compared to the troll one. On the other hand, in the first variation, a higher number of troll meme samples again introduces a class imbalance.

7. Result and Discussion

In the ResNet variations, we observed that there is no change in the macro averaged precision, recall and F1-score except for *TamilMemes + ImageNet + Flickr1k* variation. This variation has relatively poor results when compared with the other three variations in ResNet. While precision at identifying the troll class for the ResNet baseline does not vary much, we get better precision at classifying troll memes in the *TamilMemes* variation. This shows that the ResNet model trained on just Tamil memes has a better chance at identifying troll memes. The scenario is different in the case of the MobileNet variations. On the one hand, we observed less precision at identifying the troll class for the TamilMemes variation. On the other

	ResNET							
Variations	**TamilMemes**				**TamilMemes + ImageNet**			
	Precision	Recall	f1-score	count	Precision	Recall	f1-score	count
troll	**0.37**	0.33	0.35	100	**0.36**	0.35	0.35	100
not-troll	0.68	0.71	0.70	200	0.68	0.69	0.68	200
macro-avg	**0.52**	**0.52**	**0.52**	300	**0.52**	**0.52**	**0.52**	300
weighted-avg	0.58	0.59	0.58	300	0.57	0.57	0.57	300
Variations	**TamilMemes + ImageNet + Flickr1k**				**TamilMemes + ImageNet + Flickr30k**			
troll	**0.30**	0.34	0.32	100	**0.36**	0.35	0.35	100
not-troll	0.64	0.59	0.62	200	0.68	0.69	0.68	200
macro-avg	**0.47**	**0.47**	**0.47**	300	**0.52**	**0.52**	**0.52**	300
weighted-avg	0.53	0.51	0.52	300	0.57	0.57	0.57	300
	MobileNet							
Variations	**TamilMemes**				**TamilMemes + ImageNet**			
troll	**0.28**	0.27	0.28	100	**0.34**	0.43	0.38	100
not-troll	0.64	0.66	0.65	200	0.67	0.58	0.62	200
macro-avg	**0.46**	**0.46**	**0.46**	300	**0.50**	**0.51**	**0.50**	300
weighted-avg	0.52	0.53	0.52	300	0.56	0.53	0.54	300
Variations	**TamilMemes + ImageNet + Flickr1k**				**TamilMemes + ImageNet + Flickr30k**			
troll	0.33	0.55	0.41	100	0.31	0.34	0.33	100
not-troll	0.66	0.45	0.53	200	0.65	0.62	0.64	200
macro-avg	**0.50**	**0.50**	**0.47**	300	**0.48**	**0.48**	**0.48**	300
weighted-avg	0.55	0.48	0.49	300	0.54	0.53	0.53	300

Table 1: Precision, recall, F1-score and count for ResNet, MobileNet and their variations.

hand, we see improvement in precision at detecting trolls in the TamilMeme + ImageNet variation. This shows that MobileNet can leverage transfer learning to improve results. The relatively poor performance of MobileNet on the TamilMeme variation shows that it can not learn complex features like ResNet does to identify troll memes. For ResNet, the trend in the macro averaged score can be seen increasing in *TamilMemes + ImageNet* and *TamilMemes + ImageNet + Flickr1k* variations when compared to the TamilMemes variation. The *TamilMemes + ImageNet + Flickr30k* variation shows a lower macro averaged score than that of the *TamilMemes + ImageNet + Flickr1k* variation in both MobileNet and ResNet. Overall the precision for troll class identification lies in the range of 0.28 and 0.37, which is rather less than that of the not-troll class which lies in the range of 0.64 and 0.68. When we train ResNet in class imbalanced data in *TamilMemes* and *TamilMemes + ImageNet + Flickr30k* variations, results shows that the macro-averaged score of these variations are not hampered by the class imbalance issue. While for same variations MobileNet shows poor macro-averaged precision and recall score when compared with other variations. This shows that MobileNet is more susceptible to class imbalance issue than ResNet.

8. Conclusions and Future work

As shown in the Table 1 the classification model performs poorly at identifying of troll memes. We observed that this stems from the problem characteristics of memes. The meme dataset is unbalanced and memes have both image and text embedded to it with code-mixing in different forms. Therefore, it is inherently more challenging to train a classifier using just images. Further, the same image can be used with different text to mean different things, potentially making the task more complicated.

To reduce the burden placed on annotators, we plan to use a semi-supervised approach to the size of the dataset. Semi-supervised approaches have been proven to be of good use to increase the size of the datasets for under-resourced scenarios. We plan to use optical character recognizer (OCR) followed by a manual evaluation to obtain the text in the images. Since Tamil memes have code-mixing phenomenon, we plan to tackle the problem accordingly. With text identification using OCR, we will be able to approach the problem in a multi-modal way. We have created a meme dataset only for Tamil, but we plan to extend this to other languages as well.

Acknowledgments

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2, co-funded by the European Regional Development Fund, as well as by the H2020 project Prêt-à-LLOD under Grant Agreement number 825182.

Bibliographical References

Atanasov, A., De Francisci Morales, G., and Nakov, P. (2019). Predicting the role of political trolls in social media. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 1023–1034, Hong Kong, China, November. Association for Computational Linguistics.

Bishop, J. (2013). The effect of de-individuation of the internet troller on criminal procedure implementation: An

interview with a hater - proquest. *International journal of cyber criminology*, page 28–48.

Bishop, J. (2014). Dealing with internet trolling in political online communities: Towards the this is why we can't have nice things scale. *Int. J. E-Polit.*, 5(4):1–20, October.

Bloodgood, M. and Grothendieck, J. (2013). Analysis of stopping active learning based on stabilizing predictions. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 10–19, Sofia, Bulgaria, August. Association for Computational Linguistics.

Chakravarthi, B. R., Arcan, M., and McCrae, J. P. (2018). Improving Wordnets for Under-Resourced Languages Using Machine Translation. In *Proceedings of the 9th Global WordNet Conference*. The Global WordNet Conference 2018 Committee.

Chakravarthi, B. R., Arcan, M., and McCrae, J. P. (2019a). Comparison of Different Orthographies for Machine Translation of Under-Resourced Dravidian Languages. In Maria Eskevich, et al., editors, *2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 70 of *OpenAccess Series in Informatics (OASIcs)*, pages 6:1–6:14, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Chakravarthi, B. R., Arcan, M., and McCrae, J. P. (2019b). WordNet gloss translation for under-resourced languages using multilingual neural machine translation. In *Proceedings of the Second Workshop on Multilingualism at the Intersection of Knowledge Bases and Machine Translation*, pages 1–7, Dublin, Ireland, August. European Association for Machine Translation.

Chakravarthi, B. R., Priyadharshini, R., Stearns, B., Jayapal, A., S, S., Arcan, M., Zarrouk, M., and McCrae, J. P. (2019c). Multilingual multimodal machine translation for Dravidian languages utilizing phonetic transcription. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 56–63, Dublin, Ireland, August. European Association for Machine Translation.

Chakravarthi, B. R., Jose, N., Suryawanshi, S., Sherly, E., and McCrae, J. P. (2020a). A sentiment analysis dataset for code-mixed Malayalam-English. In *Proceedings of the 1st Joint Workshop of SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) (SLTU-CCURL 2020)*, Marseille, France, May. European Language Resources Association (ELRA).

Chakravarthi, B. R., Muralidaran, V., Priyadharshini, R., and McCrae, J. P. (2020b). Corpus creation for sentiment analysis in code-mixed Tamil-English text. In *Proceedings of the 1st Joint Workshop of SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) (SLTU-CCURL 2020)*, Marseille, France, May. European Language Resources Association (ELRA).

Clarke, I. and Grieve, J. (2017). Dimensions of abusive language on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 1–10, Vancouver, BC, Canada, August. Association for Computational Linguistics.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Dash, N. S., Selvraj, A., and Hussain, M. (2015). Generating translation corpora in Indic languages:cultivating bilingual texts for cross lingual fertilization. In *Proceedings of the 12th International Conference on Natural Language Processing*, pages 333–342, Trivandrum, India, December. NLP Association of India.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., and Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Trans. Interact. Intell. Syst.*, 2(3), September.

Galery, T., Charitos, E., and Tian, Y. (2018). Aggression identification and multi lingual word embeddings. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 74–79, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Gandhi, S., Kokkula, S., Chaudhuri, A., Magnani, A., Stanley, T., Ahmadi, B., Kandaswamy, V., Ovenc, O., and Mannor, S. (2019). Image matters: Detecting offensive and non-compliant content/logo in product images. *arXiv preprint arXiv:1905.02234*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778.

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., and Mishra, S. (2015). Analyzing labeled cyberbullying incidents on the instagram social network. In *International conference on social informatics*, pages 49–66. Springer.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Jose, N., Chakravarthi, B. R., Suryawanshi, S., Sherly, E., and McCrae, J. P. (2020). A survey of current datasets for code-switching research. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018). Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Kumar, R. (2019). # shutdownjnu vs# standwithjnu: A study of aggression and conflict in political debates on social media in india. *Journal of Language Aggression and Conflict*.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*,

33(1):159–174.

Malmasi, S. and Zampieri, M. (2017). Detecting hate speech in social media. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 467–472, Varna, Bulgaria, September. INCOMA Ltd.

Mathur, P., Shah, R., Sawhney, R., and Mahata, D. (2018). Detecting offensive tweets in Hindi-English code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26, Melbourne, Australia, July. Association for Computational Linguistics.

Mihaylov, T. and Nakov, P. (2016). Hunting for troll comments in news community forums. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 399–405, Berlin, Germany, August. Association for Computational Linguistics.

Mihaylov, T., Georgiev, G., and Nakov, P. (2015a). Finding opinion manipulation trolls in news community forums. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 310–314, Beijing, China, July. Association for Computational Linguistics.

Mihaylov, T., Koychev, I., Georgiev, G., and Nakov, P. (2015b). Exposing paid opinion manipulation trolls. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 443–450, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Mojica de la Vega, L. G. and Ng, V. (2018). Modeling trolling in social media conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153.

Nogueira dos Santos, C., Melnyk, I., and Padhi, I. (2018). Fighting offensive language on social media with unsupervised text style transfer. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 189–194, Melbourne, Australia, July. Association for Computational Linguistics.

Priyadharshini, R., Chakravarthi, B. R., Vegupatti, M., and McCrae, J. P. (2020). Named entity recognition for code-mixed Indian corpus using meta embedding. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*.

Rani, P., Suryawanshi, S., Goswami, K., Chakravarthi, B. R., Fransen, T., and McCrae, J. P. (2020). A comparative study of different state-of-the-art hate speech detection methods for Hindi-English code-mixed data. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, Marseille, France, May. European Language Resources Association (ELRA).

Ranjan, P., Raja, B., Priyadharshini, R., and Balabantaray, R. C. (2016). A comparative study on code-mixed data of Indian social media vs formal text. In *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, pages 608–611, Dec.

Rao, T. P. R. K. and Lalitha Devi, S. (2013). Tamil English cross lingual information retrieval. In Prasenjit Majumder, et al., editors, *Multilingual Information Access in South Asian Languages*, pages 269–279, Berlin, Heidelberg. Springer Berlin Heidelberg.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Suryawanshi, S., Chakravarthi, B. R., Arcan, M., and Buitelaar, P. (2020). Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, Marseille, France, May. European Language Resources Association (ELRA).

Wang, W. Y. and Wen, M. (2015). I can has cheezburger? a nonparanormal approach to combining textual and visual information for predicting and generating popular meme descriptions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 355–365, Denver, Colorado, May–June. Association for Computational Linguistics.

Xiang, G., Fan, B., Wang, L., Hong, J., and Rose, C. (2012). Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1980–1984.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.