Multimedia Tools and Applications (2023) 82:9137–9171
https://doi.org/10.1007/s11042-022-13796-x

1209: RECENT ADVANCES ON SOCIAL MEDIA ANALYTICS AND MULTIMEDIA SYSTEMS: ISSUES AND CHALLENGES

TrollsWithOpinion: A taxonomy and dataset for predicting domain-specific opinion manipulation in troll memes

Shardul Suryawanshi[1] · **Bharathi Raja Chakravarthi**[1] · **Mihael Arcan**[1] · **Paul Buitelaar**[1]

Received: 31 December 2020 / Revised: 10 June 2022 / Accepted: 5 September 2022 /
Published online: 22 September 2022

Abstract

Memes have become a de-facto media device in online communication. Unfortunately, memes are also used for trolling, which intends to demean, harass, or bully targeted individuals. As a result of which, the targeted individual could fall prey to opinion manipulation. Trolling via Image With Text (IWT) memes which we refer to as 'troll memes', are difficult to identify due to the multimodal (image + text) nature of such memes. However, the research into the identification and classification of troll memes with opinion manipulation remains unexplored. To bridge this research gap, we introduce a three-level taxonomy that studies the effect of trolling in domain-specific opinion manipulation. On the first level, we classify the meme as troll or not_troll. On the second level, we classify if the meme intends opinion manipulation. On the third level, if the opinion manipulation is present, then we classify the domain (political, product, other) of the opinion manipulation. To support the class definitions proposed in the taxonomy, we enhanced an existing dataset (Memotion) by annotating the data with our defined classes. This results in a dataset of 8,881 IWT memes in the English language (TrollsWithOpinion dataset) which we make available as open-source at Github(https://github.com/shardul s007/TrollOpinionMemes). We perform experiments on all three levels and present the classification report of the results using Machine Learning and state-of-the-art Deep Learning techniques. The classification report highlights the complex nature of the task since the models perform well on the first two levels. However, we see a degradation of the evaluation results on the third level of the taxonomy.

Keywords Troll memes classification · Offensive multimodal content · Corpora · Opinion manipulation

✉ Shardul Suryawanshi
shardul.suryawanshi@insight-centre.org

Extended author information available on the last page of the article.

1 Introduction

Richard Dawkins coined the word *meme* in his 1976 bestseller *The Selfish Gene* [9] as a unit of cultural transmission or a unit of imitation and replication. Because it acts like genes and can self-replicate, and mutate. On social media, memes take the form of media devices such as text, images, speech, GIFs, and video. Among these, Image with text (IWT) memes are a popular choice for internet users; they have an image with text imposed on them [11]. One might miss the true meaning of IWT memes by considering just the text or image in isolation; hence both image and text should be considered. This makes the problem "multimodal".

IWT memes can be used in trolling due to their ability to self-replicate and mutate amongst the targeted audience. "Trolling [3, 26] is the activity of posting a message via social media that tends to be offensive, provocative, or menacingly distracting and digressive or off-topic content with the intent of provoking the audience [13]." Internet users who participate in such activities, use a troll meme to spread hatred and manipulate their audience. We define a troll meme as a meme that contains:

I Offensive text and non-offensive images.
II Offensive images with non-offensive text
III Sarcastically offensive text with non-offensive images, or sarcastic image with offensive text

which is often provocative, distractive, digressive or off-topic with the intent to demean or offend particular people, group or race.

Although trolling seems to intend humour most of the time, trolls may have an agenda to influence the opinions of other internet users behind their humour. This agenda is also known as opinion manipulation [8], which can influence the opinions of an individual or a group of people positively or negatively. This is evident since troll farms influenced the outcomes of historical events such as the 2016 US presidential election and Brexit [2]. As these events shape the future of the world, we need a system that can identify trolling in the form of opinion manipulation.

In this paper, we present the TrollsWithOpinion dataset, which is designed to address trolling with domain-specific opinion manipulation problems in IWT memes. The dataset consists of around 8,881 IWT memes that we extracted from the dataset from the Memotion-analysis shared task [33]. For the meme emotion analysis [35, 38, 39], the dataset for this shared task was annotated on five different dimensions: sentiment analysis [19, 20, 22], sarcasm detection, offensive, motivation, and humour detection. We further annotated this dataset along the dimensions of trolling and opinion manipulation. For this purpose, we propose a three-level taxonomy. On the first level (Troll Level), we classify a meme into either troll or not_troll. On the second level (Opinion Manipulation Level), we identify the presence of opinion manipulation by labelling memes either as opinion_manipulation or without_opinion_manipulation. Lastly, on the third level (Domain-specific Level), we categorise memes based on the domain. We concentrated on the "product" and "political" domains with the remainder categorised as "other".

We experiment on each level with five Machine Learning (ML) based classifiers along with nine Deep Learning (DL) based classifiers and present our results. All the ML experiments are unimodal since they just use text modality from the meme. The DL experiments are multimodal since they use both the text and image modalities from the meme. We lay out all the experiments in Section 4.

Our contributions are threefold:

I We defined domain-specific opinion manipulation in troll memes
II We designed a three level taxonomy: Troll Level, Opinion Level and Domain-specific Level
III We developed the TrollsWithOpinion dataset based on the proposed taxonomy, and trained state-of-the-art ML and DL techniques on the developed dataset which we make available as open-source

The paper is organised as follows: Section 2 contains the related work around our research. Section 3 refers to the taxonomy, annotation process, inter-annotator agreement, examples, and detailed information on data statistics in terms of class distribution on each level of the taxonomy. Section 4 highlights the detailed information about the ML and DL experiments along with their hyperparameter settings. In the Section 5, we discuss the evaluation results from each baseline. Finally, in Section 6, we conclude the research along with the future direction.

2 Related work

Memes Zannettou et al. [41] defines memes based on the virality of the post; a set of posts (image, video) can be referred to as a meme if those posts share a common theme and are disseminated by a large number of users. Zannettou's (2018) social media analysis through the lens of memes gives an insight into the trend of sharing memes on mainstream (Reddit, Twitter) and fringe (4chan, gab) web communities. They proposed a pipeline based on a pHash and clustering algorithm, showing that the behaviour of sharing racist and/or political memes is more commonplace in fringe web communities. Syuntyurenko [36] suggests the possibility that a malicious bot can be disguised as a user, which can create a meme that might connect to a wider audience due to references to popular movies or other media. This might end up influencing the audience's opinion or ideological orientation, which is also known as opinion manipulation.

Trolling A significant amount of research has been done in the areas of hate speech detection [31], offensive speech detection [40], and identifying trolling [21] in text. Rather less emphasis has been put on other modalities such as image, audio or video. Tomaiuolo et al. [37] thoroughly assess the methods and challenges involved in identifying trolls. They highlight the use of post-based (based on the content in the online posts), thread-based (based on the analysis of thread of online posts), user-based (based on the overall attitude of the user) and community-based (based on the relationships of the user within the online community) methods to identify trolls. Our research is post-based which focuses on the identification of trolling (action) through the content of the online post. This method tends to be quicker, and would thus better enable preventing the harm that might be caused in comparison to the other methods mentioned.

Opinion Manipulation Trolling can lead to opinion manipulation; [25] and [2] point out the opinion manipulation caused by trolling in news and political media. Atanasov et al. [2] propose an approach to identify left, right and centre (news feed) trolls using supervised and distantly supervised methods while [25] use an approach that distinguishes trolls from non_trolls based on manual features derived from the number of comments posted, number of days in the forum, number of days with at least one comment, and number of publications commented on; both of these approaches are user-based. According to [4], political trolls

are state-sponsored agents who control a set of pseudonymous user accounts, also known as "sock puppets" [23], which dissipate misinformation [15] and propaganda with the purpose of swaying opinions, destabilising society and influencing election outcomes. All of these aforementioned techniques emphasise text-based methods to identify trolls, while we concentrate on the multimodal aspect of trolling in the form of a meme.

Our contribution To the best of our knowledge, none of these studies have investigated trolling through memes which may or may not cause opinion manipulation. The research conducted by [2, 4, 25, 36] touches upon opinion manipulation caused by trolling, but does not lay out granular classes of opinion manipulation based on a targeted population. Our research aims to bridge this gap by providing definitions for granular classes of troll or not_troll memes that may or may not cause opinion manipulation based on the following actions: Trolling with opinion manipulation, Trolling without opinion manipulation, Not-trolling with opinion manipulation, Not-trolling without opinion manipulation. Based on these actions, we categorise a meme into troll_opinion_X, not_troll_opinion_X, not_troll_without_opinion, troll_without_opinion where X represents a targeted domain that could be either political, product or other. We define each meme class in Section 3.1. We use this classification in defining a taxonomy and dataset. Moreover, we experimented with the TrollsWithOpinion dataset with state-of-the-art classifiers.

3 Dataset

We utilised data from the Memotion dataset and enhanced the annotations as per our task. The original dataset has memes in the form of an image and OCR extracted text by the use of Google vision API[1] which has been verified and manually corrected. The annotated memes fall into Humorous, Sarcasm, Offensive, Motivation classes with quantified intensity (not, slightly, mildly, very), to which a particular effect of a class is expressed, along with the overall sentiments (very negative, negative, neutral, positive, very positive). For our research, we only considered the image and text associated with the image.

3.1 Taxonomy

For this research, we defined the activity of posting messages on the internet as trolling or not-trolling depending on whether the manipulative intention is present and/or whether the manipulative effect is present. Inspired from [8], we define the following four actions based on the presence or absence of trolling and opinion manipulation:

I *Trolling with opinion manipulation:* strategy or manoeuvre that makes offensive yet digressive comments to negatively influence opinion about the targeted individual or group (political party, company, minority, gendered group, religious group, etc.).
II *Trolling without opinion manipulation:* strategy or manoeuvre that makes offensive yet digressive comments about the targeted individual or group but does not influence opinion about the individual or group.
III *Not-trolling with opinion manipulation:* strategy or manoeuvre that makes a non-offensive or informative comment that intends to positively influence opinion about the targeted individual or group.

[1] Google vision API

IV *Not-trolling without opinion manipulation:* Strategy or manoeuvre that makes a non-offensive or informative comment about the targeted individual or group but does not intend to influence opinion about the individual or group.

Figure 1 shows the three-level taxonomy of memes. On the first level or Troll Level, we define multimodal or IWT memes into either troll or not_troll classes. On the second level or Opinion Manipulation Level, we indicate if opinion manipulation is present (Opinion manipulation or Without opinion manipulation), and the third level or Domain-specific Level represents the domain (political, product, other). In this work, we concentrated on the political and product domain by marking the rest as "other".

Trolling is an action that can be predicted based on user history and the troll memes used [37]. However, this thread-based method often needs a significant amount of user data such as comments, and online activity, and hence is slow at predicting trolling. In this research, we are concerned with the identification of trolling based on troll memes (specifically, IWT memes) which is a post-based method. Troll memes are characterised by either one (image or text) or both (image and text) modalities being offensive. Based on the four actions defined above, we classified troll memes into eight classes as below:

troll_opinion_political : These memes intend trolling with opinion manipulation by making an offensive comment that negatively influences opinion about a political figure or party. The example circled Ⓐ from Fig. 2 is one such example, where a troll is criticising Donald Trump (45th president of the United States) in a demeaning manner.

not_troll_opinion_political : This meme intends not-trolling with opinion manipulation by making a non-offensive comment to positively influence opinion about a political figure or party. The example circled Ⓑ from Fig. 2 shows Barack Obama (44th president of the United States) smiling at the crowd with a caption stating the reason (killing Osama bin Laden) for not producing a birth certificate. This meme is trying to manipulate people into siding themselves with Obama since he killed a terrorist.

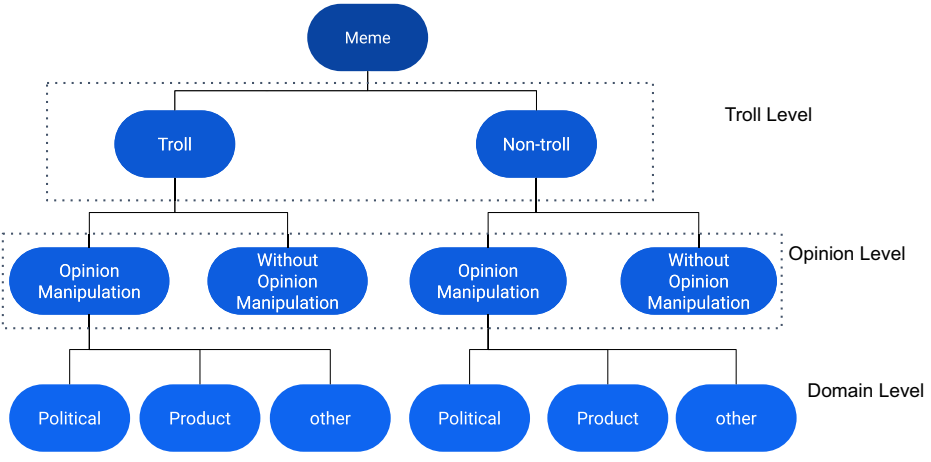


Fig. 1 Taxonomy of memes: Classification of memes based on presence of trolling along with domain-specific opinion manipulation

Fig. 2 Example of a troll and not_troll meme when political opinion manipulation is present

troll_opinion_product : This meme intends to troll with opinion manipulation by making an offensive comment to negatively influence opinion about a product (movies, video games, devices, daily household appliances, etc.) or individual that represents the product (e.g. company CEO, brand ambassador, etc.). The example circled Ⓒ from Fig. 3 is trolling Apple product users by stating the supremacy of Android users over them. The implication of the meme could be a manipulation of new mobile users, as they might get inclined towards Android devices instead of Apple devices.

not_troll_opinion_product : This meme intends not-trolling with opinion manipulation by making a non-offensive comment to positively influence opinion about a product (movies, video games, devices, daily household appliances, etc.) or individual that represents the product (e.g. company CEO, brand ambassador, etc.). The example circled Ⓓ from Fig. 3 shows a picture of Steve Jobs (ex-Apple CEO) with a motivational caption. As Steve Jobs

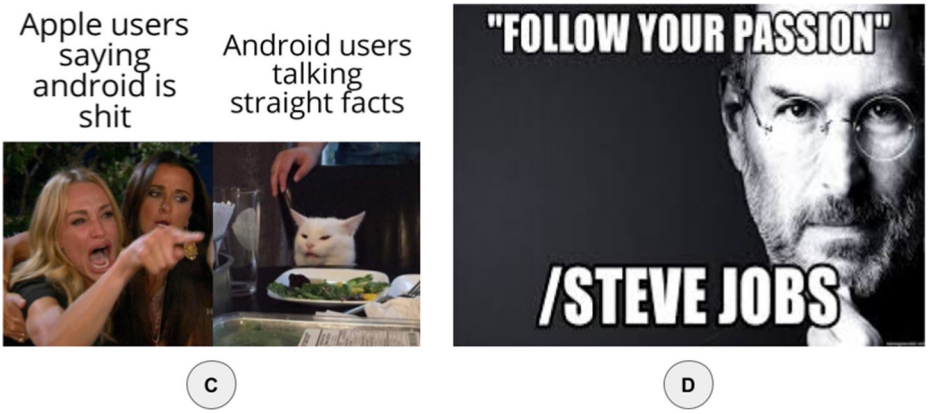


Fig. 3 Example on a troll and not_troll memes when product opinion manipulation is present

was associated with Apple, this meme is trying to win over people by putting Steve Jobs in the spotlight (associating him with a passion).

troll_opinion_other : This meme intends to troll with opinion manipulation by making an offensive comment to negatively influence opinion about an individual or group based on their gender, ethnicity, sexual orientation, religious beliefs, etc. The example circled (E) from Fig. 4 shows Selena Gomez (celebrity pop singer) with a trash bin, and the caption (Selena Gomez taking her music out for a walk) that goes along with the meme, and which is meant to criticise her music in a demeaning way.

not_troll_opinion_other : This meme intends not-trolling with opinion manipulation by making a non-offensive comment to positively influence opinion about an individual or group based on their gender, ethnicity, sexual orientation, religious beliefs, etc. The example circled (F) from Fig. 4 states that Tom Cruise (an American actor) looks young for his age by comparing him with a vampire.

troll_without_opinion : This meme intends to troll without opinion manipulation by making an offensive comment about the targeted individual or group but does not influence opinion about such individuals or groups. The example circled (G) from Fig. 5 is trolling people who play Minecraft (a graphically less intensive game) on their high-end gaming machines. But this meme is not changing opinion about "the rich kid" negatively or positively, and hence it can be classified as a troll_without_opinion category.

not_troll_without_opinion : This meme intends not-trolling without opinion manipulation by making a non-offensive comment about an individual, group or product but does not influence opinion about them. The example circled (H) from Fig. 5 is neither trolling nor changing opinion.

Fig. 4 Example of a troll and not_troll meme when opinion manipulation other than product and politics is present

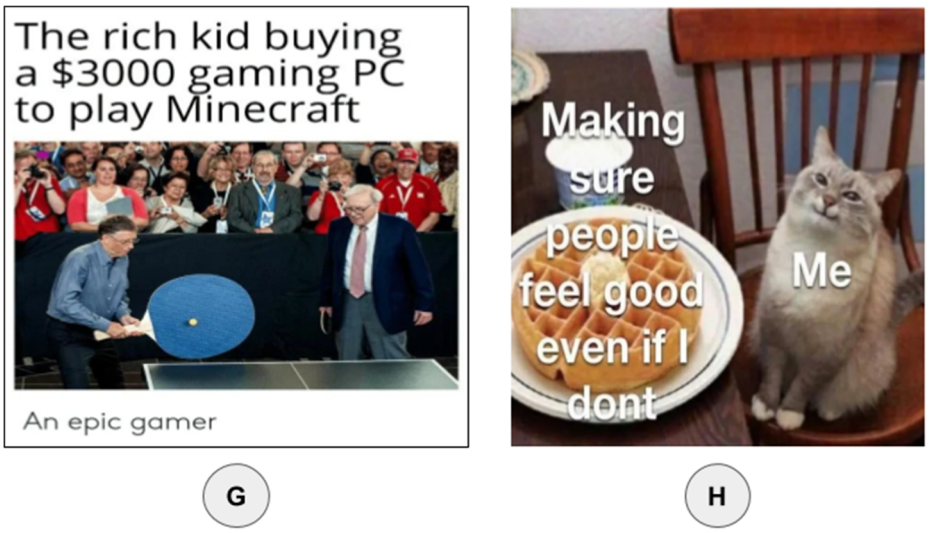


Fig. 5 Example of a troll and not_troll meme when opinion manipulation is not present

3.2 Annotation process

We did the annotation in two parts with the help of Google forms (example is provided in Fig. 8). Each Google form consisted of 50 memes, with 10 memes on each page. Annotators (volunteers) were provided with annotation guidelines, and it was made sure that they understood the class definitions provided in the annotation guidelines. However, we also modified classes and definitions as per the feedback received from annotators in the pilot annotation. To proceed with the final annotation, annotators were instructed to agree on the terms, conditions and class definitions as shown in Figs. 6 and 7. Furthermore, they were allowed to discontinue at any point if they feel overwhelmed or offended due to the content present in the forms (Fig. 8).

In the pilot annotation, we used 24 volunteers to annotate memes. The annotators were from a variety of backgrounds – in terms of gender, education, and nationality as shown in Table 1. The similarity between annotations and gold labels (annotated by first author) has been measured with the Jaccard similarity score.

Below are the findings from the pilot annotation:

I The choice or label that has been given by each annotator is subjective. This is the root cause of disagreement which is understandable as memes are vague and directed towards a particular individual or group who share similar ideas.

II The initial choices for the annotations were not_troll, troll_without_opinion, troll_opinion_x_positive, troll_opinion_x_negative and troll_opinion_x_neutral where x could be the product, political or personal. Later, we changed classes to the ones discussed in Section 3.1 because the word troll is generally negative, and cannot be associated with anything positive or neutral.

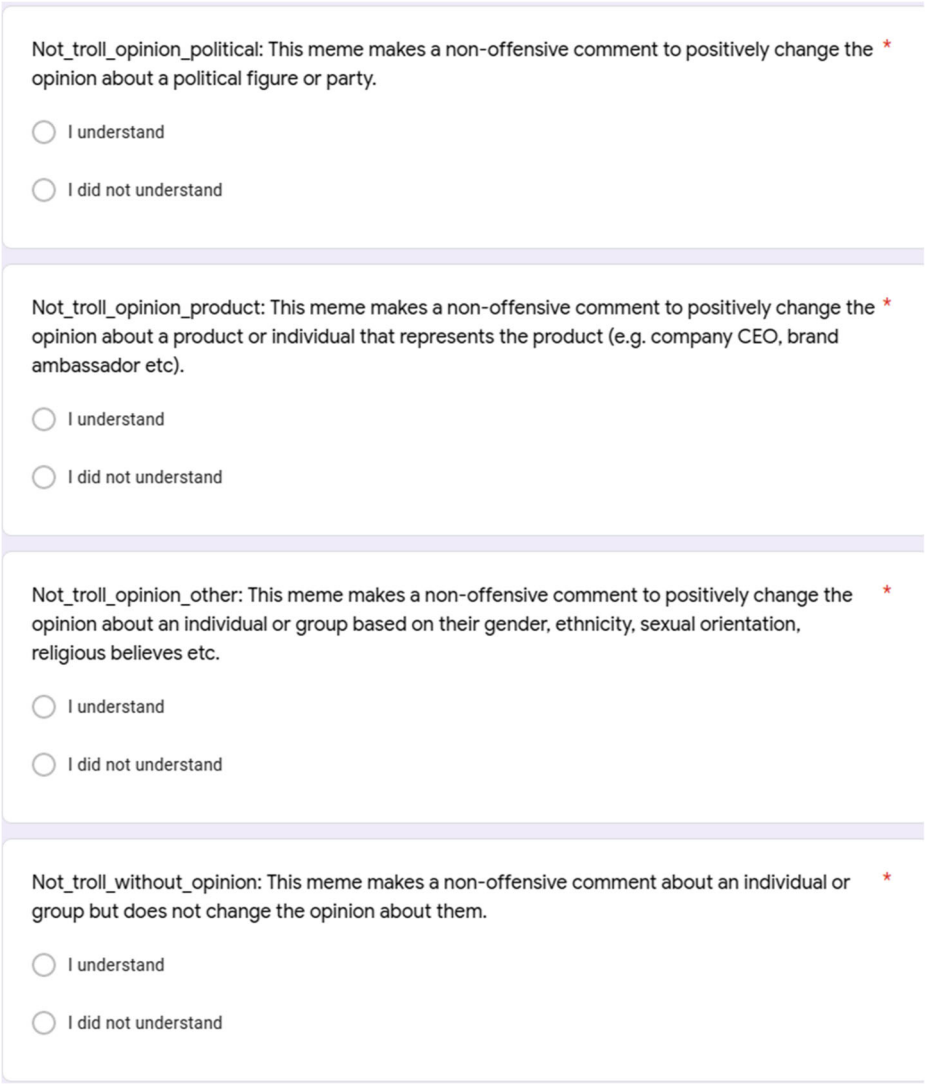


Fig. 6 A sample page from google form that explains not_troll meme classes

For the final annotation, we selected two male and two female annotators with top Jaccard similarity score[2]. Firstly, a set of ten Google forms was provided to all four annotators. Subsequently, the next set of forms was provided based on the annotators' delay in response and willingness. Each sample was annotated by at least three different annotators to establish the majority vote while deciding the final label. Hence, we needed an inter-annotator agreement that could accommodate an empty fourth annotation.

[2]Jaccard Similarity

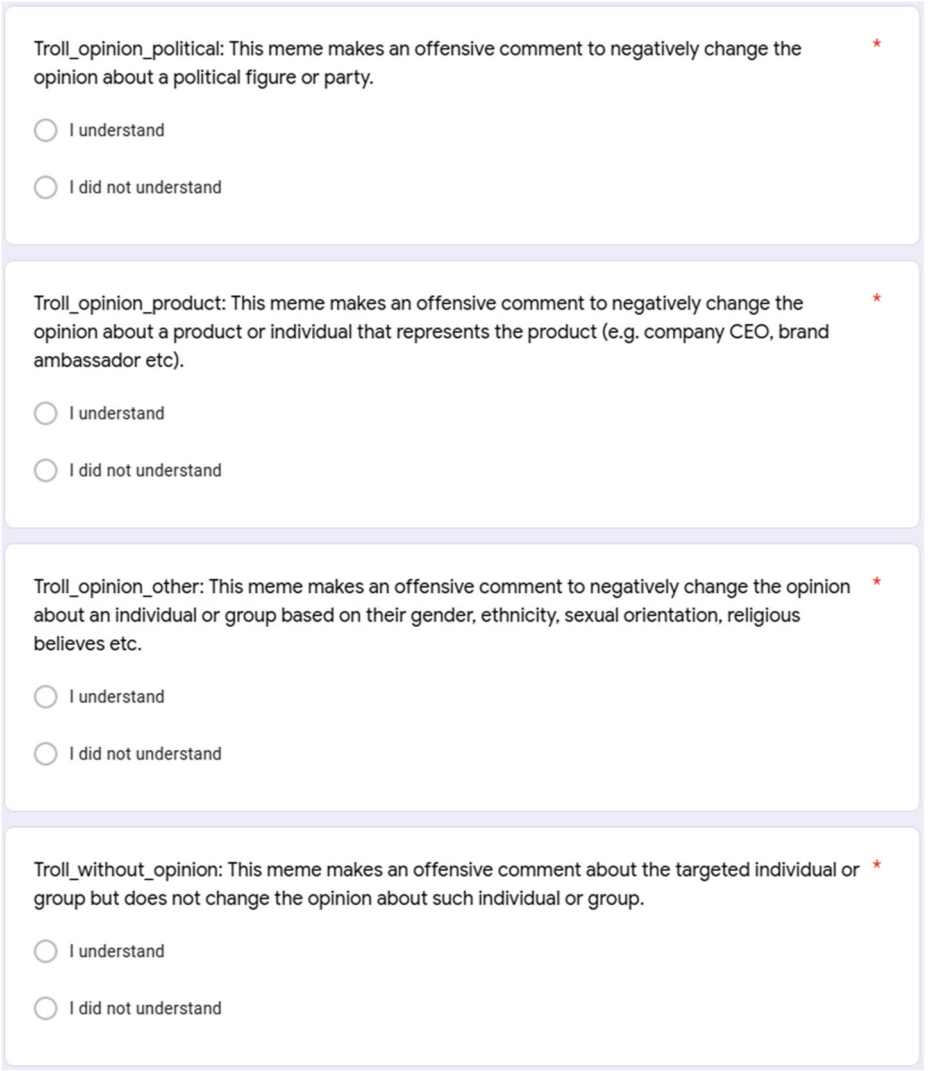


Fig. 7 A sample page from the google form that explains troll meme classes

3.3 Inter-annotator agreement

As discussed above, we used an agreement analysis that can accommodate three different annotators. In this case, Krippendorff's alpha [17] turned out to be useful. It is based on the observed disagreement corrected for disagreement expected by chance. It calculates the reliability coefficient that measures the agreement amongst annotators by distinguishing between the incomplete units (annotations) or assigning computable values to them. It is a ratio of observed disagreement D_o and expected disagreement D_e (see (1)), that ranges from -1 to 1, where 1 indicates perfect agreement, 0 indicates no agreement beyond

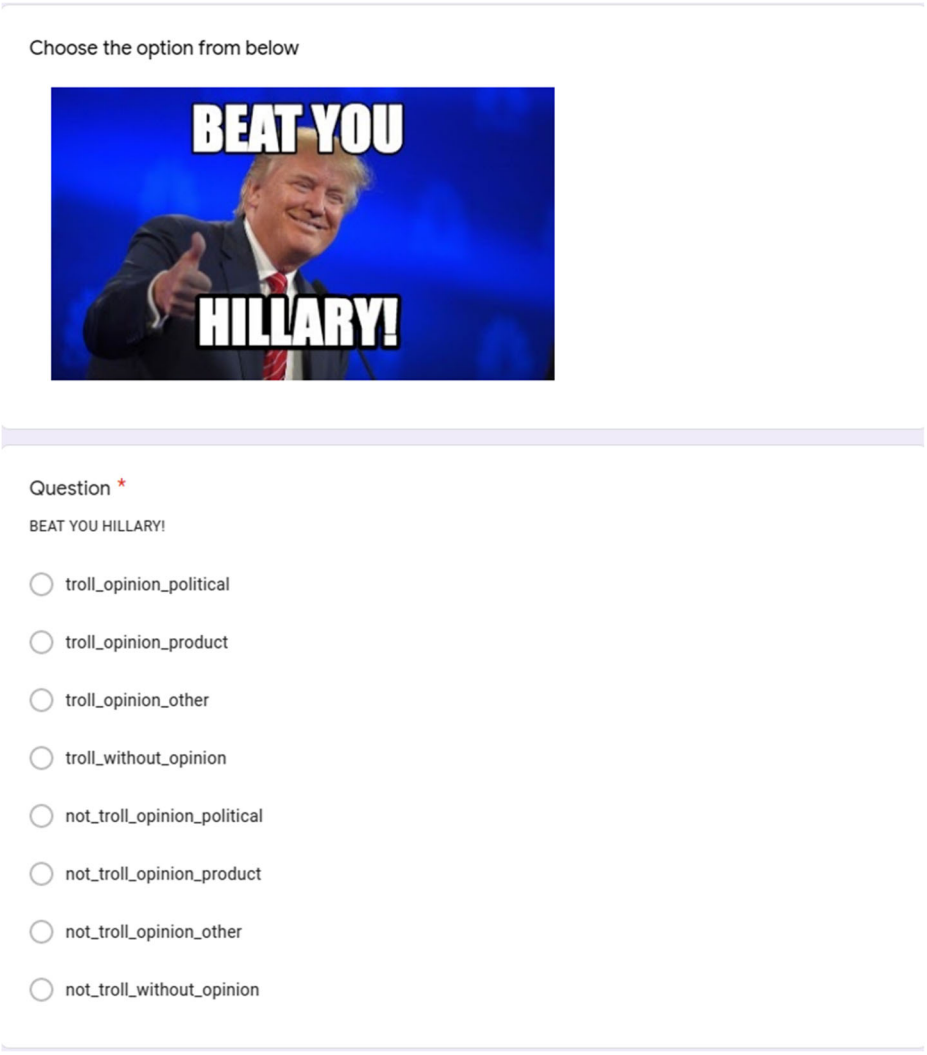


Fig. 8 An example from the Google form that provides the meme and option to choose from

chance, and negative values indicate inverse agreement.

$$\alpha = 1 - \frac{D_o}{D_e} \tag{1}$$

$$D_o = \frac{1}{n} \sum_c \sum_k o_{ck\ metric}\ \delta^2_{ck} \tag{2}$$

$$D_e = \frac{1}{n(n-1)} \sum_c \sum_k n_c\ .\ n_{k\ metric}\ \delta^2_{ck} \tag{3}$$

Here $o_{ck}\ n_c\ n_k$ and n indicates the frequencies of values in the coincidence matrices and $metric$ indicates a metric or level of measurement such as nominal, ordinal, interval, ratio

Table 1 Distribution of demographics (gender, education and age-group) for annotators in pilot annotation and final annotation

Phase		Pilot	Final
Gender	Male	14	2
	Female	9	2
	Non-binary	1	-
Education	Undergraduate	4	-
	graduate	8	2
	Postgraduate	12	2
Age group	18 to 24	4	1
	25 to 34	15	3
	35 to 44	3	-
	45 to 54	1	-
	55 to 64	1	-
Total		24	4

and others. δ^2_{ck} is a difference function which varies based on the level of measurement[3]. Krippendorff's alpha applies to all these metrics. In our case, we used the nominal and interval metric to calculate inter-annotator agreement.

We were able to achieve significant agreement amongst annotators (nominal metric: 0.699; interval metric: 0.775). The final label for each class has been decided based on the majority vote. In case of conflict, expert's (first author) vote considered as final.

3.4 Difficult examples

Memes are inherently obscure and meant for the targeted group. We observed this issue while going through annotated samples. Some samples had interesting annotations as mentioned below:

I An example circled (G) from Fig. 5 was annotated as troll_opinion_product by one annotator, as it is talking about Minecraft (product) negatively.

II An example circled (D) from Fig. 3 was annotated as not_troll_without_opinion as it is motivational if one ignores that it is associated with a company (Apple[4]).

III An example circled (H) from Fig. 5 was annotated as not_troll_opinion as it is portraying the original poster (one who posted) positively

The annotation guidelines are helpful, but it cannot be a formula or solution that fits all, and also, there is no one correct label for each meme. Hence, this task could become a multi-class, multi-label one, but we leave this to future work.

3.5 Annotators

We selected annotators from the pool of volunteers based on the Jaccard similarity index. After selecting four annotators, we performed the final annotation and selected the label of

[3] Krippendorff's alpha

[4] Apple Inc. is an American multinational technology company headquartered in Cupertino, California, that designs, develops and sells consumer electronics, computer software, and online services.

a meme based on the majority vote. In a case of conflict, we consulted with an expert for resolution.

In our pilot annotation, we saw volunteers from different demographics in terms of nationality (UK, Germany, Australia, USA, India), gender (male, female, non-binary) and education (undergraduate, graduate, postgraduate). The distribution of demographics in Table 1 shows the dominance of the age group of 25 to 34 in both phases (pilot annotation and final annotation). This involvement could be attributed to the rise of meme culture during the era of Millenials, as they are the ones who started this culture. As a result, we see less involvement from other age groups – most of them do not connect with the meme culture.

Due to resource constraints, we were not able to include people from all demographics in our pool of annotators. Hence, we relied on randomly allocated volunteers for the pilot annotation phase. This randomization in the first phase along with the Jaccard similarity index in the final annotation phase, helped us to achieve a significant agreement amongst annotators.

3.6 Data statistics

Table 2 shows an overview of the data statistics (Mean, Standard deviation, Minimum value, Maximum value) for parameters – word count, stop word count, capital word count, numerical value count and character count per sample – of the text associated with memes. On average, a meme contained 16.55 words, out of which 2.73 were stopwords. However, numerical values account for only 0.20 out of 16.55 words. Hence, we can say that stopwords[5] might hold the most valuable information, while numerical values the least. But this is based on a text-level analysis, while on a corpus-level (text from all the memes), stopwords are of negligible interest because they occur with greater frequency; thus keeping them after text processing may not be useful. Furthermore, we can see that there is a high standard deviation for each parameter. In the case of "Word count", the standard deviation of 14.61 suggests that the length of the text varies by the word count of 14 from the mean value i.e. 16.55 on average. Also, the text length could vary from a minimum of 1 to a maximum of 306. The maximum word count helps limit the length of the text. In our case, we limited the length of the text to 300 based on this statistic.

Table 3 illustrates the distribution of classes in the annotated corpus. As we are annotating an existing dataset, the distribution of the classes shows an imbalance. If memes were collected using keyword searches as per class, then a more balanced class distribution might have been achieved. However, such a balanced distribution might not represent the data in the wild (internet). While developing the Memotion dataset, the authors collected data from 52 unique and globally popular categories. This will help us to understand the distribution of proposed classes (from Table 3) in popular (majorly talked about) opinions on the Internet.

It is worth noting that the troll_without_opinion class dominates the class distribution presented in Table 3. The gap between this and the second most common class, not_troll_opinion_other, is significant. The prevalence of the troll_without_opinion might be due to the difficulty in identifying opinion manipulation, as not all meme intentions are explicit.

Figure 9 illustrates the domain (product, political, other) distribution of troll and not_troll classes irrespective of the presence of opinion manipulation. As shown, the number of troll classes – troll_opinion_X (X could be political, product, other) except

[5] nltk English stopwords

Table 2 Data statistics for the text associated with the meme from the TrollsWithOpinion dataset

Parameters (per sample)	Mean	Standard deviation	Minimum value	Maximum value
Word count	16.55	14.61	1	306
Stopword count	**2.73**	5.15	0	100
Numerical value count	0.20	0.96	0	56
Character count	91.49	76.69	2	1600

troll_without_opinion – is lower compared to that of not_troll classes – not_troll_opinion_X except not_troll_without_opinion. This disparity is due to the fact that there are fewer troll memes in popular opinions generally [41]. Figure 9 also shows that the highest number of troll memes are present in other domains, irrespective of the presence of opinion manipulation. It is not a surprise that a higher number of troll memes than not_troll memes are present in the political domain. Hence, a meme is more likely to be a troll in popular opinions if it is political. In the product domain, a higher number of not_troll than troll memes are present.

Figure 10 shows the distribution of opinion vs distribution of without_opinion for troll and not_troll classes, irrespective of domain. This graph shows that a higher number of troll memes are present in the without_opinion class when compared with the opinion class. As without_opinion class does not belong to any of the domains, it is absent from Fig. 9; however, without_opinion class is included in the opinion-wise distribution of troll and not_troll classes. We can summarise that there tends to be a higher number of not_troll memes in the opinion class, while there is a higher number of troll memes in the without_opinion class.

4 Experiments

This section describes all the experiments performed on the TrollsWithOpinion dataset. There are 14 experiments in a total of which six are based on [16], three based on the Visual Question and Answering task (VQA) (nine DL experiments) and five traditional ML experiments shown in Table 4.

The TrollsWithOpinion dataset was split into train-test-validation with 80%, 10% and 10% split ratios. All the experiments were evaluated on the held-out test set with 621 samples. We trained each experiment on every annotation level. The Troll Level classifies the given meme into the troll or not_troll category. The opinion manipulation level

Table 3 Data distribution for the classes in the TrollsWithOpinion dataset

Class	Distribution
troll_opinion_political	339
not_troll_opinion_political	249
troll_opinion_product	169
not_troll_opinion_product	815
troll_opinion_other	1,110
not_troll_opinion_other	1,622
troll_without_opinion	3,496
not_troll_without_opinion	873
Total	8,673

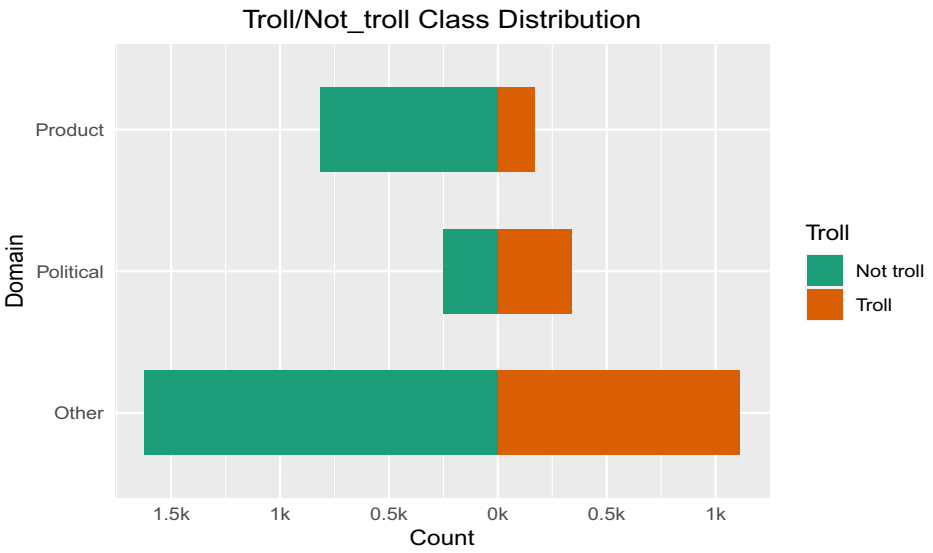


Fig. 9 Domain-specific distribution of troll and not_troll classes in the TrollsWithOpinion dataset

classifies the given meme into the opinion or without_opinion category. The Domain-specific Level classifies the given meme into any of the leave nodes categories presented in Fig. 1 i.e. troll_opinion_political, not_troll_opinion_political, troll_opinion_product, not_troll_opinion_product, troll_opinion_other, not_troll_opinion_other, troll_ without_opinion, not_troll_without_opinion. We present the results on respective levels in Section 5.

Fig. 10 Opinion-wise distribution of troll and not_troll classes the TrollsWithOpinion dataset

Table 4 List of 14 experiments performed on the TrollsWithOpinion dataset

Traditional ML Experiments	[16] Experiments	VQA Experiments
Logistic regression	Bag of words	VisualBERT
Gaussian Naive Bayes	Text-only BERT	LXMERT
Support vector machine	Image-only	UNITER
Random forest	Concat BOW + Img	
K-nearest neighbour	Concat BERT + Img	
	Multimodal Bitransformer	

4.1 Traditional ML experiments

For all the traditional ML experiments, we used state-of-the-art contextual BERT [10] embeddings. These BERT embeddings are derived from the last hidden state of DistilBERT [30]. As a result, a 768-dimensional vector representation of a text has been achieved. Later, this text vector has been used as an input to each traditional ML experiment. To ensure the optimal performance of each of these ML experiments, we incorporated a grid search to find suitable hyperparameters with the help of four-fold cross-validation [32]. L2 regularization has been used to avoid overfitting.

The general de-facto method of applying ML to the text starts with the vectorization of the text. Instead of using a tf-idf (term frequency-inverse document frequency) vector representation of the text, we used the last hidden layer representation from DistilBERT with 768 hidden units i.e. 768 dimensional (768 d) features. This contextual representation of the text acts like 768 independent features used to predict the targeted independent variable. Below, we explained each of the ML experiments. Please note that the 768 d features derived from DistilBERT were not fine-tuned, instead, we froze them.

Logistic regression (LR) is a basic building block of a neural network. We decided to leverage its simple architecture to classify a given meme based on just the text (associated with the meme). LR assumes a linear relationship between the text and a label, which predicts the best line of fit for given text features (768-dimensional vector from DistilBERT). It takes as input text features and calculates the softmax function which later has been translated into predicted classes based on the threshold value, i.e. 0.5. As we assume that these text features are linearly dependent as we do not use any kind of nonlinearity such as ReLU or GeLU. We optimized the algorithm based on the grid search over the inverse regularization strength. This experiment does not have many parameters to train on the dataset. Hence, it does not take much time to train. However, this could result in underfitting.

Gaussian Naive Bayes (GNB) [42] is a probabilistic ML experiment that naively assumes that each of the 768 features is conditionally independent of others. The Naive Bayes variation used here (GNB) assumes the underlying likelihood function to be a Gaussian distribution. Posteriori probability, i.e. probability of independent label given the 768d features, is calculated by multiplying the Gaussian likelihood function. We used the default parameter setting from the sklearn library[6] with the portion of the largest variance of all features

[6] sklearn library

set to 1e9. Since this experiment assumes that the input features are inherently independent, GNB does not consider the context represented by the DistilBERT features. However, due to this naive assumption, it would be easy to identify the emphasis of each feature while predicting the meme classes.

Support vector machine (SVM) [5] is a discriminative classification that separates data points using a hyperplane; the objective of the classifier is to keep these hyperplanes separated from each other with large margins. But the traditional version of SVM does not scale well with a large dataset, hence we are using linear SVM [12] with flexible penalties and loss functions. It uses a linear kernel, and parameters are tuned based on the hinge loss function. We trained a linear SVM with the random state of 0 (to reproduce the results) with the tolerance for stopping criteria of 1e5. This experiment relies on feature engineering, and since we are using DistilBERT features, SVM has a better chance of predicting the meme classes. However, it takes a significant amount of time to identify the vector space of each meme class while training.

Random forest (RF) [6] is an ensemble of decision tree classifiers where the data samples are distributed amongst the suitable decision trees (DT) [7] and their outcomes are averaged to improve the evaluation metric scores. Each decision tree predicts the target label by learning if-then-else decision rules. We chose RF to leverage the scaled simplicity of the DT in the form of an ensemble architecture. We trained RF with the grid search over the 50 to 200 range of the number of trees with both Gini impurity as well as entropy as a criterion to measure the quality of the split. Like SVM, this experiment also heavily on feature engineering, and similarly, the DistilBERT features could aid in RF's training. However, this ensemble experiment could take more time to train compared to the rest of the ML experiments.

K-nearest neighbour (KNN) [1] classifier is a non-generalising algorithm that assigns a class label based on the majority vote from the K-nearest neighbours. We used KNN as an experiment to classify the text associated with the meme into the required classes. We did a grid search for the value of K (number of nearest neighbours) in the [2,4,6,8]. In this experiment, neighbours are derived based on the euclidean distance between them. Being a lazy learner this experiment does not take much time in training. However, this quality may not work on a large dataset such as TrollsWithOpinion.

4.2 DL experiments

All the DL experiments are more complex than ML experiments. Unlike ML experiments, these experiments take advantage of image modality and learn the multimodal representation of the memes. Hence, they stand a better chance of identifying the underlying probability distribution of the meme classes better than ML experiments. However, they may overfit due to a significant amount of parameters.

Bag of words (BOW) This unimodal experiment uses text representation in the form of 300-dimensional pre-trained GloVe [28] (common crawl) word embeddings, which were fed to a classification layer. These GloVe embeddings were fine-tuned on word-word co-occurrence statistics from a text corpus curated out of the text from the dataset. We chose this unimodal text experiment to analyse the classification performance with the use of non-contextual word embeddings.

Text-only BERT (BERT) In this unimodal experiment, only text modality was used to train a pre-trained base uncased version of BERT to get the feature representation of the given text. This BERT representation was provided to a fully connected network, which formed a classification layer. This model is uncased as it does not make a difference between uppercase and lowercase words. Unlike sequential models, for example, Recurrent Neural Network (RNN), this transformer-based architecture does not rely on the sequence of the words. Instead, it was trained in Masked Language Modelling (MLM) fashion, which randomly masks 15% of the words from the input text, and later predicts these masked words. In this way, the inner representation (BERT features) of the text was learnt, which was later fine-tuned on the text corpus curated from our TrollsWithOpinion dataset. Unlike GloVe, BERT embeddings derived from BERT features are contextual.

Image-only (Img) This unimodal experiment makes use of image representation derived from ResNet-152 [14] pre-trained on ImageNet[7]. Later, this image representation with a 2048-dimensional feature vector was fed to a classification layer. ResNet-152 is capable of learning feature maps that are not prone to the vanishing gradient descent problem, the problem that is generally seen in the plain deep convolution networks (without residuals). The "shortcut connections" or "skip connections" proposed by this architecture perform identity mapping without adding extra parameters or complexity to the neural network.

Concat BOW + Img (ConcatBow) [16]: In this multimodal experiment, both text and image representations were used in the form of concatenated feature vectors which were derived from BOW (text) and Img (image) experiments. This concatenation of features resulted in a 2048+300-dimensional feature vector, which was fed to a classification layer. This architecture follows an early fusion technique wherein text featured from GloVe and image features from ResNet-152 were concatenated and trained jointly by feeding to a sigmoid layer.

Concat BERT + Img (ConcatBert) [16]: In this multimodal experiment, text representation from BERT and image representation from Img experiments were concatenated to form a 2048+768-dimensional feature vector. Later, this feature vector was fed to a classification layer with a sigmoid activation function (as with all experiments). The difference between ConcatBow and ConcatBert comes from textual features while the former uses GloVe (non-contextual word embedding) and the latter uses BERT (contextual word embeddings).

Multimodal Bitransformer (MMBT) [16]: This multimodal experiment used both image and text features derived from Img and BERT experiments which were later combined using bidirectional transformer architecture similar to BERT. The architecture takes in the contextual word embeddings in the form of text features derived from BERT which were added to the segment, positional embeddings. On the other hand, image features derived from ResNet were mapped to the semantic space of the contextual embeddings to form image tokens which were again added with the segment and positional embeddings before feeding to Bi-directional transformers. This architecture emphasises that the textual features carry

[7] ImageNet

more meaningful information than the image features, which is the reason behind mapping image features in the semantic space of word features.

VisualBERT This multimodal experiment pre-trained on the MS-COCO dataset [24] follows a single-stream cross-modal training approach, which enhances basic transformer blocks without making significant architectural changes. This experiment was trained on the Masked Language Modelling (MLM) objective, while it did not use Masked Vision Modelling (MVM). The original model experimented on the VQA, Visual Commonsense Reasoning (VCR), natural language for visual reasoning (NLVR) and image captioning. We finetuned the original model trained on the image caption data by removing the last layer of the model that predicts the tokens of the captions and replacing it with the classifier head with 'n' number of neurons with softmax activation function where 'n' is the number of classes to be predicted (two for Troll and Opinion Manipulation Level; eight for Domain-specific Level).

LXMERT This multimodal experiment processes the image and text with two independent, unimodal encoders. Furthermore, an additional encoder combines the two unimodal representations via cross-attention. The experiment is pre-trained on images from COCO and Visual Genome [18] as well as the image question answering datasets with the learning objective of MLM, MVM along with cross-modality matching. The original model used in the experiment was pre-trained on the image captioning data, VQA data. In our experiment, we replaced the last layer from the original model with the classifier head with 'n' neurons with a softmax activation function where 'n' is the number of classes to be predicted (two for Troll and Opinion Manipulation Level; eight for Domain-specific Level).

UNITER This multimodal experiment unlike VisualBERT leverages both MLM and MVM objectives. This transformer model utilises self-attention on the joint image and text inputs. MVM is achieved by the features extracted from Faster R-CNN [29] along with 7-dimensional location features of the bounding boxes. Similar to VisualBERT and LXMERT, UNITER is pre-trained on the MS-COCO, Visual Genome, Conceptual Captions [34], and SBU Captions [27] datasets. We finetuned the original model by replacing the last layer of the model with a fully connected network of 'n' neurons with the softmax activation function where 'n' is the number of classes to be predicted (two for Troll and Opinion Manipulation Level; eight for Domain-specific Level).

4.3 Hyperparameter settings

Table 5 shows the hyper-parameter settings for all nine DL experiments. All the experiments use an AdamW optimizer with the variable number of epochs, batch size, learning rate and maximum sequence length (max_seq_len) for the text. If the learning stagnated (no improvement in validation accuracy), then the learning rate was dropped by a factor of 0.5 of the previous learning rate.

5 Results and discussion

In this section, we discuss classification results on each level i.e. Troll Level, Opinion Manipulation Level and Domain-specific Level of the annotation scheme.

Table 5 Hyper-parameter settings for all DL experiments

Hyper-parameters	BOW	BERT	Img
Optimizer	AdamW	AdamW	AdamW
Epoch	20	20	20
Batch size	16	16	16
Learning rate	2e-5	4e-8	1e-7
Dropout	0.1	0.1	0.1
Patience	5	5	5
max_seq_len	100	100	-
Hyper-parameters	**ConcatBow**	**ConcatBert**	**MMBT**
Optimizer	AdamW	AdamW	AdamW
Epoch	20	20	20
Batch size	16	16	16
Learning rate	2e-5	4e-7	2e-5
Dropout	0.1	0.1	0.1
Patience	5	5	5
max_seq_len	30	100	30
Hyper-parameters	**VisualBERT**	**LXMERT**	**UNITER**
Optimizer	AdamW	AdamW	AdamW
Epoch	10	10	10
Batch size	18	18	18
Learning rate	1e-6	1e-6	1e-6
Dropout	-	-	-
Patience	-	-	-
max_seq_len	62	62	62

5.1 Evaluation results at the troll level

In this task, we trained both ML and DL experiments on the binary supervised data with the samples either labelled as troll or not_troll. The supervised data has a class imbalance with 5,114 samples labelled as "troll" and the rest i.e. 3,559 as "not_troll". Table 6 shows the evaluation result of ML experiments in terms of precision, recall and f-score. In the ML experiments, LR was able to identify the troll memes at the precision of 64% which stays at the top when compared with other ML experiments' precision. As LR has been trained on the text modality, its highest precision score emphasises the importance of the text modality over the image as the evaluation scores (Precision of troll: 57%, Precision of not_troll: 45%) for the Img experiment are lower than that of LR. The recall of the troll class for the RF experiment is highest with 80%. Moreover, the RF experiment has the highest f-score i.e. 70% for the troll class. This shows that the decision tree-based experiment could effectively identify troll memes based on text modality. Tables 7, 8 and 9 show the evaluation results of DL experiments in terms of precision, recall and f-score respectively.

Amongst all the DL experiments, the maximum precision of the troll class is 64% for the MMBT experiment, while the maximum recall is 94% for ConcatBert. Amongst all the multimodal experiments (ConcatBow, ConcatBert, MMBT, VisualBERT, LXMERT, UNITER), MMBT shows a balanced performance in terms of all evaluation metrics (Precision for troll:64%, Recall for the troll: 77% and f-score for the troll: 70%). The balanced performance by MMBT could be attributed to the inclusion of both text and image modalities in

Table 6 Precision, Recall, and f-score for Machine Learning experiments trained on the TrollsWithOpinion dataset

	LR	SVM	GNB	RF	KNN	Count
	Precision					
troll	0.64	0.60	0.61	0.62	0.58	350
not_troll	0.53	0.49	0.51	0.58	0.47	271
Weighted avg	0.59	0.56	0.57	0.60	0.54	621
	Recall					
troll	0.63	0.64	0.67	0.80	0.66	350
not_troll	0.54	0.46	0.44	0.35	0.38	271
Weighted avg	0.59	0.56	0.57	0.61	0.56	621
	f-score					
troll	0.64	0.62	0.64	0.70	0.62	350
not_troll	0.54	0.48	0.48	0.44	0.42	271
Weighted avg	0.59	0.56	0.57	0.59	0.53	621

BERT-like architecture that combines image feature maps from ResNet and text features from the BERT using self-attention and multi-head attention under the MLM objective. In Table 7, improvement in precision of both troll as well as not_troll classes could be seen in the multimodal counterparts of BOW (Precision for troll: 57%, Precision for not_troll 44%) and BERT (Precision for troll: 58%, Precision for not_troll 44%) which are Concat-Bow (Precision for troll: 60%, Precision for not_troll 57%) and ConcatBert (Precision for troll: 58%, Precision for not_troll 59%). However, Table 8 shows the improvement in the recall of troll classes for the multimodal counterparts but the recall of the not_troll class has been reduced. Hence, it could be deduced that the inclusion of image modality will improve the precision while the recall shows an unpredictable trend, that may or may not improve. Figure 11 is a bar graph comparing the weighted average precision, recall and f-score of ML experiments at Troll Level. Here, it could be seen that the maximum weighted average Precision, recall and f-score are 62%, 62% and 61% in the multimodal MMBT experiment.

Table 7 Precision at Troll Level for Deep Learning experiments trained on the TrollsWithOpinion dataset

	Precision		
	Img	**BOW**	**ConcatBow**
troll	0.57	0.57	0.60
not_troll	0.45	0.44	0.57
Weighted avg	0.52	0.52	0.58
	Bert	**ConcatBert**	**MMBT**
troll	0.58	0.58	0.64
not_troll	0.44	0.59	0.59
Weighted avg	0.52	0.58	0.62
	VisualBERT	**LXMERT**	**UNITER**
troll	0.53	0.55	0.55
not_troll	0.60	0.60	0.60
Weighted avg	0.57	0.58	0.58

Table 8 Recall at Troll Level for Deep Learning experiments trained on the TrollsWithOpinion dataset

	Recall		
	Img	**BOW**	**ConcatBow**
troll	0.61	0.44	0.84
not_troll	0.42	0.57	0.27
Weighted avg	0.52	0.52	0.59
	Bert	**ConcatBert**	**MMBT**
troll	0.23	0.94	0.77
not_troll	0.78	0.12	0.44
Weighted avg	0.52	0.58	0.62
	VisualBERT	**LXMERT**	**UNITER**
troll	0.36	0.32	0.27
not_troll	0.75	0.79	0.83
Weighted avg	0.58	0.59	0.59

5.2 Evaluation results at the opinion manipulation level

In this task, both ML and DL experiments are trained on the binary supervised dataset where each sample is either labelled as the opinion (total samples: 3,489) or without_opinion (total samples: 5,184). Tables 11, 12 and 13 show the evaluation results of ML and DL experiments on the binary supervised dataset in terms of precision, recall and f-score respectively. Table 10 shows that the maximum precision (66%), recall (61%) and f-score (60%) at correctly classifying into opinion class has been obtained by the RF, KNN and KNN amongst all the ML experiments. Table 11 shows that the lowest precision (48%) is seen for the Img experiment, which uses only the image modality for classification. A significant jump in the precision of ConcatBow (Precision for opinion: 61%, Precision for without_opinion: 52%) and ConcatBert (Precision for opinion: 60%, Precision for without_opinion: 49%) could be seen in the same table over their unimodal counterparts BOW (Precision for opinion: 52%, Precision for without_opinion: 50%) and BERT (Precision for opinion: 55%, Precision for without_opinion: 0%). Here, the inclusion of the image resulted in an improvement

Table 9 f-score at Troll Level for Deep Learning experiments trained on the TrollsWithOpinion dataset

	f-score		
	Img	**BOW**	**ConcatBow**
troll	0.59	0.50	0.70
not_troll	0.44	0.50	0.36
Weighted avg	0.52	0.47	0.55
	Bert	**ConcatBert**	**MMBT**
troll	0.33	0.72	0.70
not_troll	0.56	0.20	0.51
Weighted avg	0.43	0.49	0.61
	VisualBERT	**LXMERT**	**UNITER**
troll	0.43	0.41	0.36
not_troll	0.67	0.68	0.69
Weighted avg	0.56	0.56	0.55

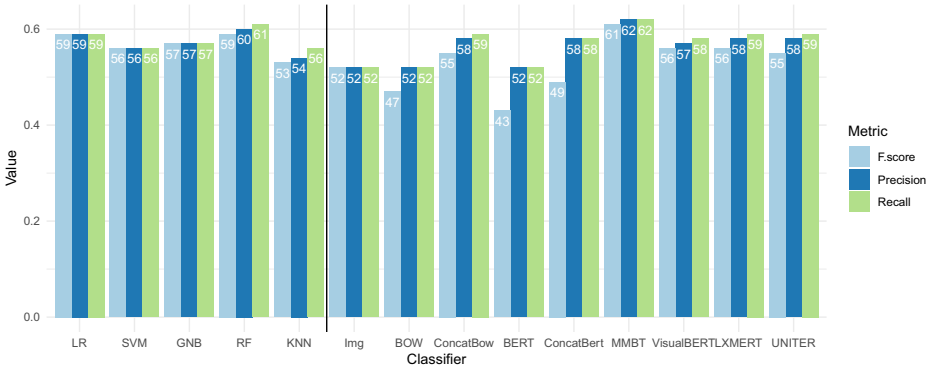


Fig. 11 A bargraph for comparison of weighted-average precision, recall and f-score across ML (left hand side) and DL (right hand side) experiments at Troll Level

Table 10 Precision, Recall, and f-score at Opinion Manipulation Level for Machine Learning experiments trained on the TrollsWithOpinion dataset

	LR	SVM	GNB	RF	KNN	Count
	Precision					
opinion	0.66	0.58	0.61	0.64	0.59	350
without_opinion	0.54	0.48	0.49	0.52	0.50	271
Weighted avg	0.61	0.53	0.56	0.59	0.55	621
	Recall					
opinion	0.53	0.55	0.44	0.50	0.61	350
without_opinion	0.54	0.51	0.66	0.66	0.48	271
Weighted avg	0.59	0.53	0.54	0.57	0.55	621
	f-score					
opinion	0.59	0.56	0.51	0.56	0.60	350
without_opinion	0.60	0.49	0.56	0.58	0.49	271
Weighted avg	0.59	0.56	0.53	0.57	0.55	621

Table 11 Precision at Opinion Manipulation Level for Deep Learning experiments trained on the TrollsWithOpinion dataset

	Precision		
	Img	**BOW**	**ConcatBow**
opinion	0.48	0.52	0.61
without_opinion	0.41	0.41	0.52
Weighted avg	0.55	0.47	0.57
	Bert	**ConcatBert**	**MMBT**
opinion	0.55	0.60	0.69
without_opinion	0.00	0.49	0.50
Weighted avg	0.30	0.55	0.61
	VisualBERT	**LXMERT**	**UNITER**
opinion	0.52	0.54	0.53
without_opinion	0.66	0.62	0.67
Weighted avg	0.60	0.58	0.61

Table 12 Recall at Opinion Manipulation Level for Deep Learning experiments trained on the TrollsWithOpinion dataset

	Recall		
	Img	**BOW**	**ConcatBow**
opinion	0.45	0.50	0.59
without_opinion	0.42	0.57	0.55
Weighted avg	0.53	0.47	0.57
	Bert	**ConcatBert**	**MMBT**
opinion	1.00	0.50	0.33
without_opinion	0.00	0.59	0.82
Weighted avg	0.55	0.54	0.55
	VisualBERT	**LXMERT**	**UNITER**
opinion	0.70	0.54	0.69
without_opinion	0.48	0.62	0.51
Weighted avg	0.58	0.58	0.59

Table 13 f-score at Opinion Manipulation Level for Deep Learning experiments trained on the TrollsWithOpinion dataset

	f-score		
	Img	**BOW**	**ConcatBow**
opinion	0.55	0.51	0.60
without_opinion	0.51	0.42	0.52
Weighted avg	0.53	0.47	0.57
	Bert	**ConcatBert**	**MMBT**
opinion	0.71	0.54	0.45
without_opinion	0.00	0.54	0.62
Weighted avg	0.39	0.54	0.53
	VisualBERT	**LXMERT**	**UNITER**
opinion	0.60	0.54	0.60
without_opinion	0.55	0.62	0.58
Weighted avg	0.57	0.58	0.59

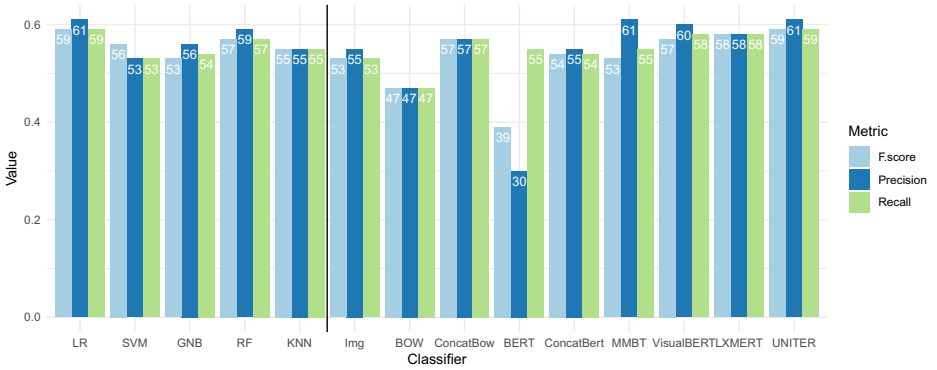


Fig. 12 A bargraph for comparison of weighted-average precision, recall and f-score across ML (left hand side) and DL (right hand side) experiments at Opinion Manipulation Level

Table 14 Precision, Recall, and f-score at Domain-specific Level for Machine Learning experiments trained on the TrollsWithOpinion dataset

	LR	SVM	GNB	RF	KNN	Count
	Precision					
troll_opinion_political	0.16	0.07	0.09	0.25	0.20	25
not_troll_opinion_political	0.17	0.15	0.14	0.36	0.25	22
troll_opinion_product	0.13	0.14	0.17	**0.67**	**1.00**	23
not_troll_opinion_product	0.18	0.11	0.12	0.19	0.20	57
troll_opinion_other	0.19	0.26	0.19	0.27	0.27	94
not_troll_opinion_other	0.40	0.32	0.32	0.38	0.28	121
troll_without_opinion	**0.44**	**0.41**	**0.42**	0.37	0.35	208
not_troll_without_opinion	0.24	0.15	0.18	0.32	0.10	71
Weighted avg	0.31	0.28	0.28	0.34	0.29	621
	Recall					
troll_opinion_political	0.32	0.08	0.12	0.16	0.12	25
not_troll_opinion_political	0.41	0.27	0.32	0.18	0.09	22
troll_opinion_product	**0.43**	0.17	0.22	0.09	0.04	23
not_troll_opinion_product	0.28	0.11	0.18	0.11	0.11	57
troll_opinion_other	0.17	0.29	0.14	0.06	0.13	94
not_troll_opinion_other	0.22	0.24	0.14	0.20	0.26	121
troll_without_opinion	0.20	**0.36**	**0.38**	**0.81**	**0.65**	208
not_troll_without_opinion	0.35	0.20	0.31	0.10	0.03	71
Weighted avg	0.24	0.26	0.25	0.36	0.31	621
	f-score					
troll_opinion_political	0.21	0.07	0.10	**0.20**	0.15	25
not_troll_opinion_political	0.24	0.19	**0.20**	0.24	0.13	22
troll_opinion_product	0.20	0.15	0.19	0.15	0.08	23
not_troll_opinion_product	0.22	0.11	0.14	0.14	0.14	57
troll_opinion_other	0.18	**0.27**	0.16	0.10	0.17	94
not_troll_opinion_other	**0.29**	0.27	**0.20**	0.26	0.27	121
troll_without_opinion	0.27	0.38	**0.40**	**0.51**	**0.45**	208
not_troll_without_opinion	**0.29**	0.17	0.23	0.15	0.04	71
Weighted avg	0.25	0.27	0.25	0.29	0.26	621

in precision. However, when compared with the precision of the opinion class by LR, DL experiments are proven inferior. Hence, it could be said that complex DL experiments with a significant amount of parameters may not be more effective than simple ML experiments. The MMBT experiment, which showed promising evaluation results at the Troll Level failed to perform at the Opinion Manipulation Level, due to poor recall (33%) for the opinion class. However, it still holds rank 1 at precision (69%) for the opinion class. This inconsistency in the performance by MMBT shows that no single classifier will be perfect for the two different tasks, proving the no free lunch theorem right. Evaluation results from Tables 11, 12 and 13 shows that all the multimodal experiments have better precision, recall

Table 15 Precision at Domain-specific Level for Deep Learning experiments trained on the TrollsWithOpinion dataset

	Precision		
	Img	**BOW**	**ConcatBow**
troll_opinion_political	0.05	0.00	0.00
not_troll_opinion_political	0.02	0.00	0.10
troll_opinion_product	0.00	0.25	0.00
not_troll_opinion_product	0.10	0.09	**0.40**
troll_opinion_other	0.17	0.20	0.33
not_troll_opinion_other	0.24	0.20	0.28
troll_without_opinion	**0.34**	**0.29**	0.36
not_troll_without_opinion	0.23	0.05	0.25
Weighted avg	0.22	0.19	0.29
	Bert	**ConcatBert**	**MMBT**
troll_opinion_political	0.03	0.00	0.00
not_troll_opinion_political	0.00	0.00	0.00
troll_opinion_product	0.02	0.00	0.00
not_troll_opinion_product	0.09	**1.00**	0.00
troll_opinion_other	0.15	**1.00**	**0.50**
not_troll_opinion_other	0.25	0.15	0.29
troll_without_opinion	**0.31**	0.00	0.36
not_troll_without_opinion	0.00	0.34	0.22
Weighted avg	0.19	0.39	0.28
	VisualBERT	**LXMERT**	**UNITER**
troll_opinion_political	0.00	0.00	0.00
not_troll_opinion_political	0.00	0.00	0.00
troll_opinion_product	0.00	0.00	0.00
not_troll_opinion_product	0.33	0.37	**0.75**
troll_opinion_other	0.00	0.00	0.00
not_troll_opinion_other	0.29	0.26	0.14
troll_without_opinion	0.35	0.36	0.34
not_troll_without_opinion	**0.71**	**0.43**	0.40
Weighted avg	0.29	0.25	0.26

and f-score than the unimodal experiments. Hence, this shows again that the inclusion of the image improves precision in identifying the opinion class. Table 13 shows that the UNITER experiments highest weighted average f-score (59%) for the opinion class. Figure 12 is a bar graph comparing the weighted average precision, recall and f-score of ML experiments at the Opinion Manipulation Level. Here, it can be seen that the maximum weighted average Precision, recall and f-score are 61%, 59% and 59% by the unimodal LR and multimodal UNITER experiments. Hence, we can see that the text modality plays an important role while identifying opinion manipulation in memes.

5.3 Evaluation results at the domain-specific level

In this multi-class classification task, troll and not_troll meme classes that are branched to opinion and without_opinion classes are divided based on the domain (political, product, other). The corresponding annotation scheme resulted in eight classes. All the DL and ML experiments are evaluated on the given supervised multi-class dataset, and results are presented in Tables 14, 15 and 16.

Table 14 shows the performance of the ML experiments on the given multi-class dataset. It can be seen that the evaluation scores for the troll_without_opinion class are the highest, irrespective of the experiment. On the other hand, troll _opinion_political showed poor evaluation scores across all the experiments with the maximum precision, recall and f-score values at 16%, 32% and 21%. The weighted average scores (precision: 34%, recall: 36%, f-score: 29%) for the RF experiment are the highest among all the ML experiments. Tables 15 and 16 show the evaluation results of DL experiments in terms of precision, recall and f-score respectively. Amongst all the DL experiments ConcatBow demonstrates the highest weighted average precision (39%) while the LXMERT showed the highest recall (35%) and f-score (23%). All the DL experiments performed poorly at classification for all the classes except troll_without_opinion and not_troll_opinion_other. Figure 13 is a bar graph comparing the weighted average precision, recall and f-score of ML experiments at the Opinion Manipulation Level. Here, it can be seen that the maximum weighted average Precision, recall and f-score are 39%, 35% and 35%. Overall, due to the lack of data for each class, the performance of both ML and DL experiments is compromised when compared to evaluation results at the Troll and Opinion Manipulation Levels.

5.4 Qualitative analysis

In this section, we will go through the test sample predictions for all the classes on Troll Level, Opinion Manipulation Level, and Domain-specific Level. For simplicity, we present the predicted labels from the classifiers which have shown the highest weighted precision at each level.

Troll Level Table 17 shows the qualitative analysis of predictions at the Troll Level by the experiment (MMBT) with the highest weighted f-score on the test data. Here, Fig. 14a is a troll meme, that has also been correctly predicted as a troll meme by MMBT. This meme demeans a political figure (President of the North Korea) by body shaming. It is a difficult example since the meaning of the meme can be understood only when we consider both the image and text modality of the meme. Moreover, Fig. 14c which is a not_troll has been also predicted as not_troll by MMBT. Unlike the previous example, this meme has a lot of text from the movie (The Matrix) reference. This meme criticises the history that has been taught in the classroom. Moreover, there is no specific target that is being demeaned via this meme. Similarly, a text-heavy meme from Fig. 14e has been correctly identified as a troll meme. Hence, the text modality plays a vital role in predicting troll memes. On the other hand, memes from Fig. 14b, d, f and h been wrongly predicted by MMBT. In the case of the meme from Fig. 14b, MMBT fails, as it incorrectly predicted the meme as a not_troll meme. But this meme is targeted to demean a product (Facebook) which could also be understood just based on the text modality. Interestingly, a meme from Fig. 14h is labelled as not_troll while MMBT predicted it as a troll meme. Even for humans, this meme could be defined as a troll, since it is demeaning to readers based on their age. But in our case, the annotators agreed to label the meme as not_troll since they consider this a joke. This again proves the

Table 16 Recall and f-score at Domain-specific Level for Deep Learning experiments trained on the TrollsWithOpinion dataset

	Recall		
	Img	**BOW**	**ConcatBow**
troll_opinion_political	0.16	0.00	0.00
not_troll_opinion_political	0.05	0.00	0.09
troll_opinion_product	0.00	0.09	0.00
not_troll_opinion_product	0.04	0.04	0.07
troll_opinion_other	0.04	0.21	0.01
not_troll_opinion_other	0.35	**0.36**	0.14
troll_without_opinion	**0.42**	0.33	**0.87**
not_troll_without_opinion	0.07	0.01	0.06
Weighted avg	0.23	0.22	0.34
	Bert	**ConcatBert**	**MMBT**
troll_opinion_political	0.08	0.00	0.00
not_troll_opinion_political	0.00	0.00	0.00
troll_opinion_product	0.04	0.00	0.00
not_troll_opinion_product	0.04	0.02	0.00
troll_opinion_other	0.15	0.01	0.03
not_troll_opinion_other	0.01	0.02	0.08
troll_without_opinion	**0.55**	0.00	**0.93**
not_troll_without_opinion	0.00	**0.99**	0.08
Weighted avg	0.22	0.34	0.34
	VisualBERT	**LXMERT**	**UNITER**
troll_opinion_political	0.00	0.00	0.00
not_troll_opinion_political	0.00	0.00	0.00
troll_opinion_product	0.00	0.00	0.00
not_troll_opinion_product	0.04	0.12	0.05
troll_opinion_other	0.00	0.00	0.00
not_troll_opinion_other	0.12	0.09	0.03
troll_without_opinion	**0.94**	**0.95**	**0.96**
not_troll_without_opinion	0.07	0.04	0.03
Weighted avg	0.35	0.35	0.33
		f-score	
	Img	**BOW**	**ConcatBow**
troll_opinion_political	0.08	0.00	0.00
not_troll_opinion_political	0.03	0.00	0.09
troll_opinion_product	0.00	0.13	0.00
not_troll_opinion_product	0.05	0.05	0.12
troll_opinion_other	0.07	0.21	0.02
not_troll_opinion_other	0.28	0.26	0.19
troll_without_opinion	**0.38**	**0.31**	**0.50**
not_troll_without_opinion	0.11	0.02	0.09
Weighted avg	0.21	0.20	0.23

Table 16 (continued)

	Recall		
	Bert	**ConcatBert**	**MMBT**
troll_opinion_political	0.04	0.00	0.00
not_troll_opinion_political	0.00	0.00	0.00
troll_opinion_product	0.02	0.00	0.00
not_troll_opinion_product	0.05	0.03	0.00
troll_opinion_other	0.15	0.02	0.06
not_troll_opinion_other	0.02	0.03	0.13
troll_without_opinion	**0.40**	0.00	**0.52**
not_troll_without_opinion	0.00	**0.51**	0.12
Weighted avg	0.17	0.18	0.22
	VisualBERT	**LXMERT**	**UNITER**
troll_opinion_political	0.00	0.00	0.00
not_troll_opinion_political	0.00	0.00	0.00
troll_opinion_product	0.00	0.00	0.00
not_troll_opinion_product	0.06	0.18	0.1
troll_opinion_other	0.00	0.00	0.00
not_troll_opinion_other	0.17	0.13	0.05
troll_without_opinion	**0.51**	**0.52**	**0.50**
not_troll_without_opinion	0.13	0.08	0.05
Weighted avg	0.22	0.23	0.19

fact that troll memes are subjective. Identifying such boundary line memes as troll meme is rather safer since the audience exposed is not known.

Opinion Manipulation Level Table 18 shows the qualitative analysis of predictions at the Opinion Manipulation Level by the experiment (MMBT) with the highest weighted average f-score on the test data. Memes from Fig. 14b and f have been correctly predicted by MMBT,

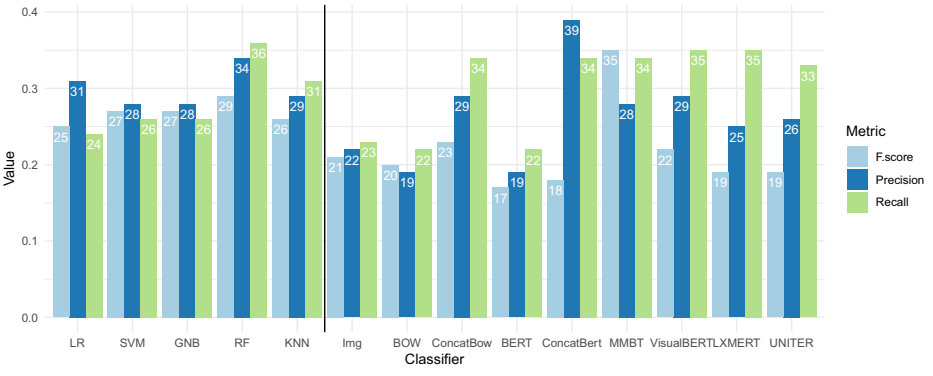


Fig. 13 A bargraph for comparison of weighted-average precision, recall and f-score across ML (left hand side) and DL (right hand side) experiments at Domain-specific Level

Table 17 Qualitative Analysis at Troll Level to analyse the predictions of the experiment with highest weighted average f-score on the test data (image column refers to images from Fig. 14)

Troll Level		
Image	True Label	MMBT Prediction (61 %)
Figure 14a	Troll	Troll
Figure 14c	not_troll	not_troll
Figure 14b	Troll	not_troll
Figure 14d	not_troll	Troll
Figure 14e	Troll	Troll
Figure 14f	not_troll	Troll
Figure 14g	Troll	Troll
Figure 14h	not_troll	Troll

while the rest of the memes from the same figure have been incorrectly predicted. Specifically, the meme from Fig. 14b is an opinion manipulation meme, where the online user (who posted the meme) intended to change the opinion of the exposed audience about the product (Facebook). Similarly, MMBT succeeds in predicting meme from Fig. 14f although an opinion manipulation meme. As both the memes can be understood based on the text modality, these memes are relatively simple to classify compared to the meme from Fig. 14a which requires an understanding of both the modalities. Moreover, the meme from Fig. 14g has been correctly identified as a without_opinion meme. However, MMBT failed at correctly predicting the opinion manipulation memes from Fig. 14e and c as they are text-heavy and clear in their intention i.e. opinion manipulation. The first one calls out on the body plus sized people and the second one frowns upon the internet audience over the topic of history.

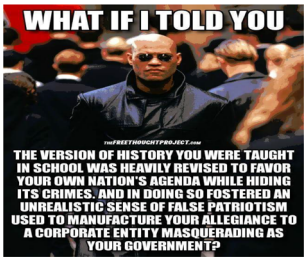
Domain-specific Level Table 19 shows the qualitative analysis of predictions at the Domain-specific Level by the experiment (UNITER) with the highest weighted average f-score on the test data. Overall, the effect of the class imbalance shown in Fig. 10 can be seen in the predicted labels i.e. more than 50% of the training data comprises samples from the troll_without_opinion class which was identified as most of the times, as shown in Table 19. This effect is prominently observed even after using class weights. Hence, we could say that enough examples from the rest of the classes could result in improvement in the evaluation metric (precision, recall, f-score). From the table, it can be seen that the UNITER was able to correctly classify memes from Fig. 14g, e, d and h. Moreover, the experiment was able to classify memes at the Troll Level and Opinion Manipulation Level. For example, even though Fig. 14b has been predicted as Fig. 14e, the classifier correctly predicted the presence of trolling and opinion manipulation but with incorrect domain i.e. product. However, the classifier incorrectly predicted the label for Fig. 14f on both Troll and Opinion Manipulation Levels, as the true label was not_troll_opinion_other but the predicted label was troll_without_opinion.

6 Conclusion

Our work introduced a meme dataset, TrollsWithOpinion, with a taxonomy that studies the effect of troll memes in the form of domain-specific opinion manipulation. We introduced a taxonomy that enhanced the Memotion dataset to form the TrollsWithOpinion dataset. We

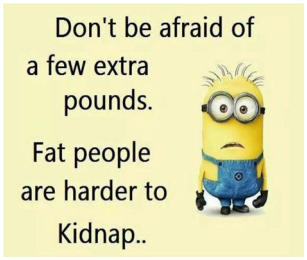
(a) troll_opinion_political

(b) troll_opinion_product

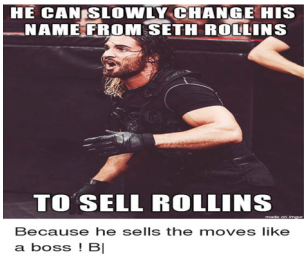


(c) not_troll_opinion_political

(d) not_troll_opinion_product

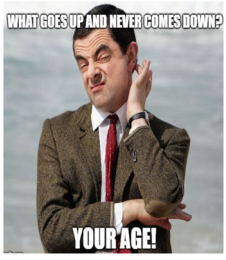


(e) troll_opinion_other



(f) not_troll_opinion_other

(g) troll_without_opinion



(h) not_troll_without_opinion

Fig. 14 Samples from each class from the Memotion test set

think that it is important to understand the effect of troll or not_troll memes that might result in opinion manipulation. Although it is hard to identify troll memes with opinion manipulation, we hope that the TrollsWithOpinion dataset will facilitate research in multimodal troll meme classification.

Table 18 Qualitative Analysis at Opinion Manipulation Level to analyse the predictions of the experiment with highest weighted average f-score on the test data (image column refers to images from Fig. 14)

Opinion Manipulation Level		
image	True Label	MMBT Prediction (61%)
Figure 14a	opinion	without_opinion
Figure 14c	opinion	without_opinion
Figure 14b	opinion	opinion
Figure 14d	opinion	without_opinion
Figure 14e	opinion	without_opinion
Figure 14f	opinion	opinion
Figure 14g	without_opinion	without_opinion
Figure 14h	without_opinion	opinion

Below are the critical findings derived from the Results and Discussion:

I **Textual features play a major role in the classification** of multimodal memes since all the text-only experiments showed a balanced performance on every level.

II **Multimodal experiments showed contradictory trends**; ConcatBow and ConcatBert showed a gain in precision in Troll as well as Opinion Manipulation Levels, while they suffered from poor recall.

III **Multimodal DL experiments showed poor evaluation metric scores compared to ML experiments in certain cases.** For example, on the Opinion Manipulation Level, we have seen that ML experiments performed better than DL experiments.

IV Also, we saw **poor performances in DL experiments on the Domain-specific Level**. This trend points to the need for more data, since a complex DL model requires more data to tune its hyperparameters. Hence, including more training data will be one of the future directions.

Hence, including more training data will be one of the future directions. Moreover, because of data imbalances, our experiments could not generalise classes with the least training samples even after introducing class weights – undersampling classes with the majority, and oversampling ones in the minority. Hence, we need a better approach or algorithm which will improve the evaluation metric scores despite the data imbalance. Also, one more way to mitigate against this can be collecting more annotations, which can potentially turn this task into a multi-label, multi-class classification problem.

Table 19 Qualitative Analysis at Domain-specific Level to analyse the predictions of the experiment with highest weighted average f-score on the test data (image column refers to images from Fig. 14)

Domain-specific Level		
image	True Label	MMBT Prediction (61%)
Figure 14a	troll_opinion_political	troll_without_opinion
Figure 14c	not_troll_opinion_political	troll_opinion_other
Figure 14b	troll_opinion_product	troll_opinion_other
Figure 14d	not_troll_opinion_product	not_troll_opinion_product
Figure 14e	troll_opinion_other	troll_opinion_other
Figure 14f	not_troll_opinion_other	troll_without_opinion
Figure 14g	troll_without_opinion	troll_without_opinion
Figure 14h	not_troll_without_opinion	not_troll_without_opinion

Acknowledgements 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 for the Insight SFI Research Centre for Data Analytics, co-funded by the European Regional Development Fund. We would like to thank all the annotators for their valuable efforts throughout the development of the dataset.

Data Availability The datasets generated during and analysed during the current study are available in the **TrollOpinionMemes** repository.

Declarations

Conflict of Interests The authors have no relevant financial or non-financial interests to disclose.

References

1. Altman NS (1992) An introduction to kernel and nearest-neighbor nonparametric regression. Amer Statistic 46(3):175–185
2. Atanasov A, De Francisci Morales G, Nakov P (2019) Predicting the role of political trolls in social media. In: Proceedings of the 23rd conference on computational natural language learning (CoNLL). Association for Computational Linguistics, Hong Kong, pp 1023–1034. https://doi.org/10.18653/v1/K19-1096. https://www.aclweb.org/anthology/K19-1096
3. Bishop J (2014) Representations of 'trolls' in mass media communication: a review of media-texts and moral panics relating to 'internet trolling'. Int J Web Based Commun 10(1):7–24
4. Boatwright BC, Linvill DL, Warren PL (2018) Troll factories:The internet research agency and state-sponsored agenda building. Resource Centre on Media Freedom in Europe
5. Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on computational learning theory, pp 144–152
6. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
7. Breiman L, Friedman J, Stone CJ, Olshen RA (1984) Classification and regression trees. CRC press
8. Coxall M (2013) Human Manipulation-A Handbook. Malcolm Coxall-Cornelio Books
9. Dawkins R (1976) The selfish gene. Oxford University Press, New York
10. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:181004805
11. Du Y, Masood MA, Joseph K (2020) Understanding visual memes: an empirical analysis of text superimposed on memes shared on twitter. In: Proc Int AAAI Conf Web Soc Media, vol 14, pp 153–164
12. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ (2008) Liblinear: a library for large linear classification. J Mach Learn Res 9:1871–1874
13. Hardaker C (2010) Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. J Politeness Res 6(2):215–242
14. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
15. Karadzhov G, Gencheva P, Nakov P, Koychev I (2018) We built a fake news & click-bait filter: what happened next will blow your mind! arXiv:180303786
16. Kiela D, Bhooshan S, Firooz H, Testuggine D (2019) Supervised multimodal bitransformers for classifying images and text. arXiv:190902950
17. Krippendorff K (2011) Computing krippendorff's alpha-reliability. Computing 1:25–2011
18. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li LJ, Shamma DA et al (2017) Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int J Comput Vision 123(1):32–73
19. Kumar K, Kurhekar M (2017) Sentimentalizer: Docker container utility over cloud. In: 2017 ninth international conference on advances in pattern recognition (ICAPR), pp 1–6. https://doi.org/10.1109/ICAPR.2017.8593104
20. Kumar K, Bamrara R, Gupta P, Singh N (2020a) M2p2: Movie'S trailer reviews based movie popularity prediction system Pant M, Sharma TK, Verma OP, Singla R, Sikanderm A (eds), Springer, Singapore
21. Kumar R, Ojha AK, Lahiri B, Zampieri M, Malmasi S, Murdock V, Kadar D (eds) (2020b) Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying. European Language Resources Association (ELRA), Marseille. https://www.aclweb.org/anthology/2020.trac-1.0 11-16 May 2020

22. Kumar S, Kumar K (2018) Irsc: Integrated automated review mining system using virtual machines in cloud environment. In: 2018 Conference on Information and Communication Technology (CICT), pp 1–6. https://doi.org/10.1109/INFOCOMTECH.2018.8722387
23. Kumar S, Cheng J, Leskovec J, Subrahmanian V (2017) An army of me: Sockpuppets in online discussion communities. In: Proceedings of the 26th international conference on World Wide Web, pp 857–866
24. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: Common objects in context. In: European conference on computer vision. Springer, pp 740–755
25. Mihaylov T, Georgiev G, Nakov P (2015) Finding opinion manipulation trolls in news community forums. In: Proceedings of the nineteenth conference on computational natural language learning. Association for Computational Linguistics, China, pp 310–314. https://doi.org/10.18653/v1/K15-1032. https://www.aclweb.org/anthology/K15-1032
26. Mojica de la Vega LG, Ng V (2018) Modeling trolling in social media conversations. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki. https://aclanthology.org/L18-1585 7–12 May 2018
27. Ordonez V, Kulkarni G, Berg T (2011) Im2text: Describing images using 1 million captioned photographs. In: Shawe-Taylor J, Zemel R, Bartlett P, Pereira F, Weinberger KQ (eds), vol 24. Advances in Neural Information Processing Systems, Curran Associates Inc. https://proceedings.neurips.cc/paper/2011/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf 7-12 May 2018
28. Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
29. Ren S, He K, Girshick R, Sun J (2016) Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39(6):1137–1149
30. Sanh V, Debut L, Chaumond J, Wolf T (2019) Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv:191001108
31. Schmidt A, Wiegand M (2017) A survey on hate speech detection using natural language processing. In: Proceedings of the Fifth international workshop on natural language processing for social media. Association for Computational Linguistics, Valencia, pp 1–10. https://doi.org/10.18653/v1/W17-1101. https://www.aclweb.org/anthology/W17-1101
32. Shao J (1993) Linear model selection by cross-validation. J Amer Stat Assoc 88(422):486–494
33. Sharma C, Scott Paka W, Bhageria D, Das A, Poria S, Chakraborty T, Gambäck B (2020) Task Report: Memotion Analysis 1.0 @SemEval 2020: The Visuo-Lingual Metaphor! In: Proceedings of the 14th international workshop on semantic evaluation (SemEval-2020). Association for Computational Linguistics, Spain
34. Sharma P, Ding N, Goodman S, Soricut R (2018) Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Melbourne, pp 2556–2565. https://doi.org/10.18653/v1/P18-1238. https://aclanthology.org/P18-1238
35. Sharma S, Kumar P, Kumar K (2017) Lexer: Lexicon based emotion analyzer. In: Shankar BU, Ghosh K, Mandal DP, Ray SS, Zhang D, Pal SK (eds) Pattern recognition and machine intelligence. Springer International Publishing, Cham, pp 373–379
36. Syuntyurenko O (2015) Network technologies for information warfare and manipulation of public opinion. Sci Tech Inf Process 42(4):205–210
37. Tomaiuolo M, Lombardo G, Mordonini M, Cagnoni S, Poggi A (2020) A survey on troll detection. Fut Int 12(2):31. 10 Feb 2020
38. Vijayvergia A, Kumar K (2018) Star: rating of reviews by exploiting variation in emotions using transfer learning framework. In: 2018 Conference on information and communication technology (CICT), pp 1–6. https://doi.org/10.1109/INFOCOMTECH.2018.8722356
39. Vijayvergia A, Kumar K (2021) Selective shallow models strength integration for emotion detection using glove and lstm. Multimed Tools Appl 80(18):28349–28363
40. Zampieri M, Malmasi S, Nakov P, Rosenthal S, Farra N, Kumar R (2019) SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In: Proceedings of the 13th international workshop on semantic evaluation. Association for Computational Linguistics, Minneapolis, pp 75–86. https://doi.org/10.18653/v1/S19-2010. https://www.aclweb.org/anthology/S19-2010
41. Zannettou S, Caulfield T, Blackburn J, De Cristofaro E, Sirivianos M, Stringhini G, Suarez-Tangil G (2018) On the origins of memes by means of fringe web communities. In: Proceedings of the internet measurement conference, vol 2018, pp 188–202
42. Zhang H (2005) Exploring conditions for the optimality of naive bayes. Int J Pattern Recognit Artif Intell 19(02):183–198

Affiliations

Shardul Suryawanshi[1] · Bharathi Raja Chakravarthi[1] · Mihael Arcan[1] · Paul Buitelaar[1]

Bharathi Raja Chakravarthi
bharathi.raja@insight-centre.org

Mihael Arcan
mihael.arcan@insight-centre.org

Paul Buitelaar
paul.buitelaar@insight-centre.org

[1] Insight SFI Research Centre for Data Analytics, Data Science Institute, National University of Ireland Galway, Galway, Ireland