Content Moderation Across Multiple Platforms with Capsule Networks and Co-Training

Student Name: Vani Agarwal

IIIT-D-MTech-CS-GEN-19-MT17068

May, 2019

Indraprastha Institute of Information Technology
New Delhi

Thesis Committee
Dr. Arun Balaji (Chair)
Dr. Ponnurangam Kumaraguru (Co-chair)
Dr. Rajiv Ratn Shah
Dr. Niharika Sachdeva

Submitted in partial fulfillment of the requirements
for the Degree of M.Tech. in Computer Science,
in General Category

Keywords: Content Moderation, Capsule Network, Co-training

Certificate

This is to certify that the thesis titled **"Content Moderation Across Multiple Platforms with Capsule Networks and Co-Training "** submitted by **Vani Agarwal** for the partial fulfillment of the requirements for the degree of *Master of Technology* in *Computer Science & Engineering* is a record of the bonafide work carried out by her under our guidance and supervision in the Security and Privacy group at Indraprastha Institute of Information Technology, Delhi. This work has not been submitted anywhere else for the reward of any other degree.

Dr. Arun Balaji & Dr. Ponnurangam Kumaraguru

Indraprastha Institute of Information Technology, New Delhi

Abstract

Social media systems provide a platform for users to freely express their thoughts and opinions. Although this property represents incredible and unique communication opportunities, it also brings along important challenges. Often, content which constitutes hate speech, abuse, harmful intent proliferates online platforms. Since problematic content reduces the health of a platform and negatively affects user experience, communities have terms of usage or community norms in place, which when violated by a user, leads to moderation action on that user by the platform. Unfortunately, the scale at which these platforms operate makes manual content moderation near impossible, leading to the need for automated or semi-automated content moderation systems. For understanding the prevalence and impact of such content, there are multiple methods including supervised machine learning and deep learning models. Despite the vast interest in the theme and wide popularity of some methods, it is unclear which model is most suitable for a certain platform since there have been few benchmarking efforts for moderated content. To that end, we compare existing approaches used for automatic moderation of multimodal content on five online platforms: Twitter, Reddit, Wikipedia, Quora, Whisper. In addition to investigating existing approaches, we propose a novel Capsule Network based method that performs better due to its ability to understand hierarchical patterns. In practical scenarios, labeling large scale data for training new models for a different domain or platform is a cumbersome task. Therefore we enrich our existing pre-trained model with a minimal number of labeled examples from a different domain to create a co-trained model for the new domain. We perform a cross-platform analysis using different models to identify which model is better. Finally, we analyze all methods, both qualitatively and quantitatively, to gain a deeper understanding of model performance, concluding that our method shows an increase of 10% in average precision. We also find that the co-trained models perform well despite having less training data and may be considered a cost-effective solution.

Acknowledgments

It is my privilege to express my sincerest gratitude to my advisors, Dr. Ponnurangam Kumaraguru and Dr. Arun Balaji, for giving me this opportunity to work on this thesis. I would also like to thank them for their valuable inputs, guidance, encouragement and wholehearted support throughout the thesis. I would like to thank my esteemed committee members, Dr. Rajiv Ratn Shah and Dr. Niharika Sachdeva for agreeing to evaluate my thesis work. I am also grateful to all the members of my Precog family at IIIT Delhi who have consistently helped me with their inputs and suggestions on the work, especially Indira Sen for shepherding me and spending her valuable time in coming up with this thesis. Special thanks to Snehal Gupta and Asmit Kumar Singh for their inputs.

Last but not the least, I would like to thank all my supportive family and friends who encouraged me and kept me motivated throughout the thesis.

Contents

List of Figures

List of Tables

Chapter 1

Research Motivation and Aim

1.1 Research Motivation

Social media sites are platforms that showcase user-generated content to engage participants. These participants are provided an abundance of reach, freedom to express their opinions and receive feedback at marginal cost. This online content covers each and every minute detail of a user on a daily basis. Despite the huge benefits of social media, it also brings along unique challenges. Often, users encounter issues like cyberbullying, online threats, abuse, harassment and hate speech. There has been an enormous increase in objectionable content on different social media platforms [39]. Be it any website related to news, business or cultural events, negative content is rising. This daily onslaught of disturbing posts can lead to conflicts on the web. Such content needs to be pulled down from websites. Measures must be taken to reduce this content. This is where content moderation comes into the picture. It is an important and relevant problem as many platforms are struggling to solve it. The article Punishing Ecstacy of being a Reddit moderator [1], highlights the variation in content posted by different users. Thus, content moderation needs to be embedded as a crucial part, for the maintenance of any social media platform.

Before the advent of social media, news related to events was more localised, and it took more time to spread. However, with widespread usage of social media, information spreads like wildfire and targets a larger set of audience making social media as a powerful tool for spreading information [18]. A study reveals that when people hear information, they're likely to remember only 10 percent of that information for three days [32]. However, if a relevant image is paired with that same information, people can retain 65 percent of the information for three days. Referring to social media, Facebook posts with images achieve 2.3 times more engagement than those without images and Buffer reported that for its user base, tweets with images received 150 percent more retweets than tweets without images [32].

With the rise of users on different social media platforms, the use of abusive language in comments or posts has become a ubiquitous part of the online social media(OSM). Every platform has certain policies to curb/reduce sensitive content. Is the content of a particular website safe for children to view? The work in [41] analyses the problems faced by children and alerts the society about the issues. Different social media platforms have different norms. So, can a model trained on one platform work on another platform, either *as-is* or with minor modifications? The majority of the existing approaches create a new dataset and proposes a model to classify the content, in essence re-inventing the wheel for every platform. Despite the awareness about the problem and the existence of various methods to solve the problem, it is not clear which method works the best on all the platforms.

As we can see, Twitter has content related to sexism and racism, Quora contains insincere questions containing questions not seeking valid answers, Wikipedia has content containing abusive words on wikipedia talk page, Whisper and Reddit also have abusive words. HateSpeech is a subset of content that needs to be moderated.

So there is a need to analyze the performance of different existing models and make up for the weaknesses uncovered. Given how important and widespread this problem is, there is a need to understand how different properties of the content affect automated content moderation techniques.

Capsule Networks learn the vector representation of the data rather than scalar values, so it captures semantic similarity among different words better. Co-training [22] is used when small amount of training data and large amount of unlabelled data is available. It is used to check the interpretability of different methods on different datasets by labelling the unlabelled data. After labelling top few annotations can be checked manually to ensure correctness of labels.

Figure 1.1: Example of sensitive post on Twitter. It shows the use of abusive words in the post.

1.2 Research Aim

Sensitive content is a problem for almost every platform. Can a single model be used to tackle the problem on different platforms like Reddit or Wikipedia which have divergent community norms? If not, then it is essential to investigate the reasons for variable performance and propose changes that can be made to the model to improve performance on other platforms. Furthermore, comparisons between existing automated content moderation techniques can help us understand the limitations of existing methods and identify gaps.

Given posts P = p1,p2,. . . ., pk from domains D = D1,D2,. . . ., Dn,
find a subset of posts which should be flagged for moderation.

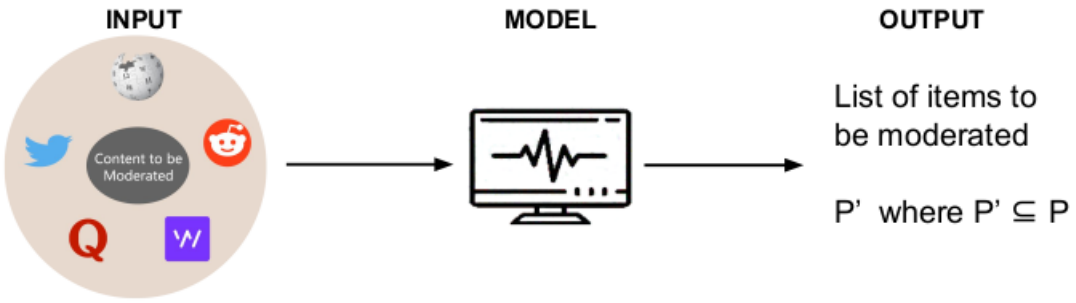

Figure 1.2: Diagrammatic explanation of overall thesis concept.

1.3 What is Content Moderation?

Different platforms have certain policies, terms of use and guidelines. Content posted on any platform which does not adhere to these guidelines need to be moderated or removed from platform. It is necessary as it causes harmful repercussions, reduces the overall efficacy of the platform as well as the user experience. The article [19] shows that people react less strongly to malicious speech on online platforms and see the victim less harmed than if words were said directly to a person. Another study [15] which shows that 8 out of 10 Indians have faced online harassment.

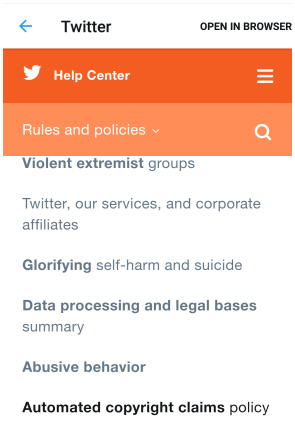

Figure 1.3: Example of policies on Twitter. It shows the use of abusive words, glorifying words, violent content is prohibited on Twitter.

Figure 1.3 shows rules and guidelines on Twitter. The rules says that violent extremist content, abusive language, and other rules are objectionable. As we can see in figure 1.1, there are abusive words like f**k, a** etc. So this post will be flagged for moderation. Similarly different platforms have different policies and posts not abiding the policies will be flagged for moderation.

Chapter 2

Related Work

Our work is an extension of the work done by Gupta et al. [28] where they propose an architecture for multimodal content moderation on Twitter. The authors focused on events related to crisis, violence, and protests i.e. content which is important for first responders such as law enforcement agencies. They explained various approaches for both text and image classification. Our work aims to delve deeper by analyzing different social media platforms data using existing techniques and seek solutions for cross-platform content moderation. We also propose a new architecture that outperforms the models proposed in the paper.

2.1 Content Moderation

User engagement on different websites, blogs, forums, and social media is increasing at a rapid pace [17]. Higher user engagement result in more comments, reviews, and likes. Not all user-generated content adheres to the norms of the particular platform. Examples of problematic content include negative opinions, unverified information, and claims [1]. Such content compromises the reputation of the platform, negatively effects user experience and can have negative offline repercussions. A study about "Content Moderation needs in 2019" [2] highlights the need of Artificial Intelligence(AI) related technological solutions rather than manually tagging content. Further, tagging of sub-categories becomes a lot more challenging when humans do it and they are prone to more errors than AI solutions [2]. Manual tagging also affects the mental health of content moderators and can lead to symptoms of PTSD [16].

Reddit has been used as a platform for hate speech and the article "You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech" [3] highlights the violence created on this platform. This violence led to a shutdown of 2278 subreddits with millions of members in a single day [23, 35]. As stated in this article [1], users post explicit content in subreddits which are not related the content's ideology, which further causes an increase of hate content. Platforms

[1] https://www.theverge.com/2019/5/25/18639754/facebook-nancy-pelosi-video-fake-clip-distorted-deepfake

combat such instances with policies concerning anti-social behavior. In response to subreddits being abused, the site introduced anti-harassment policies [4]. Kumar et al. [31] conclude that posts are initiated by active members of the community and carried away by the less active members i.e. even the less active members start commenting by seeing any sensitive post. The authors further state that, not just intra-community interactions but inter-community(among members of different subreddits) interactions also lead to the start of conflicts.This results in high inter-community conflicts even after having dedicated subreddits for different topics.

There are Wikipedia editors who edit content on Wikipedia and can also converse among themselves. One editor may ask another editor who previously edited the content the basis for an edit. This conversation leads to the use of harsh language and adversarial interactions. To combat heated arguments, Wikipedia has a policy of- Do not make personal attacks anywhere in Wikipedia [20]. Ellery et al. [45] analyze personal attacks on Wikipedia using crowdsourcing and machine learning techniques adding to the literature on understanding content moderation. Joni et al. worked on identifying and classifying hate on Facebook and Youtube and used multiple machine learning models such as Support Vector Machines(SVM), Logistic regression(LR), etc. on TF-IDF features to classify content [38]. Mondal et al. [34] highlight the abundance of hate speech on Twitter and Whisper. Their work mainly focuses on identifying the most common hate expressions on the two different platforms. The authors also study the effect of anonymity on hate speech and the most hated groups. They emphasize the need for comparison of different forms of hate on different social media platforms. Comparison of methods is especially important because existing methods do not take into account data from more than two platforms. Our work attempts to bridge this gap by studying the efficacy of content moderation algorithms in multiple platforms.

2.2 Textual approaches

Waseem et al. [44] provide a dataset of Twitter posts related to hate speech. The posts are filtered by identifying the linguistic features which contribute to racism and sexism related hate. Davidson et al. [25] categorize tweets into three categories - hate speech, offensive language and neither. They show that racism-related tweets are predicted as hate speech and sexism related tweets as part of offensive language. Wanh et al. [43] investigate the role of anonymity and lack of persistent social links on user engagement on Whisper. Badjatiya et al. [21] present a deep learning model and test it on a benchmark Twitter hate speech dataset. They extract features using basic feature extraction techniques like TF-IDF, word2vec and advanced deep learning techniques like Long short term memory(LSTM) and convolutional neural networks(CNN). These features are fed to classifiers like logistic regression and SVM. We have also used this dataset referred later as Twitter2.

Georga et al [27] use CNN for toxic comment classification on Wikipedia. Whereas, Plattner

et al. [36] show that the ensemble of Logistic regression and neural network can boost classifier performance. They also show that augmentation of hand-crafted features with word and character level n-grams boosts the classifier performance. Srivastava et al. [42] gave used Capsule Networks recently to identify aggression in wikipedia comments. They have shown a significant increase in output results using capsule networks.

2.3 Multimodal approaches

Gupta et al. [28] have used deep learning approaches to extract features from image and text data. After training individual models for image and text, each, the output feature vector is fused to predict whether the tweet needs a moderator's attention or not. Mohan et al [33] investigate how user engagement on the Reddit affects the toxicity level, discovering that subreddits of different sizes have different levels of toxicity.
Silva et al [40] provide a characterization study focused on quantitatively identifying the main targets of hate speech on Twitter and Whisper. They studied different hate speech forms and identified important patterns. This provides an elaborate understanding of different hate speech forms and also offers directions to develop prevention and detection approaches. Previous work on Reddit multimodal content by Hessel et al [29] highlights the need for multimodal content analysis due to an increasing percentage of image content on social media. They analyze and predict the relative popularity of the two items posted at approximately the same time on Reddit. Non-content features like the popularity of the person who posted the post, time at which the post is posted, weekend vs weekdays, etc. also play a major role in the popularity of image than only content features- image and text.

User interests and platform capabilities vary, therefore one platform cannot cater to all the needs of a user. Hence, we do a comparative analysis of five platforms to identify content that requires moderation.

Chapter 3

Contributions

As established in the previous section, there is a need for automated content moderation solutions to reduce the burden on human moderators. Previous research has investigated multiple methods for different platforms, but to the best of my knowledge, there has been no effort to understand the applicability of a particular content moderation method across different platforms. To that end,

- We perform **Multi-platform comparison** for Content Moderation
  checking across 5 platforms: Twitter, Reddit, Quora, Wikipedia, Whisper
  as well as checking across three baseline methods [28, 44, 45]. A single method is not best for all platforms.
- We also investigate the efficacy of **Capsule Networks** for Content Moderation
  Capsule networks are an innovative deep learning method which exploits hierarchical patterns to model complex relationships that may manifest in content that should be moderated. On an average Capsule Network performed better than LSTM by 10% in average precision.
- Finally, we leverage a **Co-training for Domain Adaptation** method [24] to analyze cross-platform performance, commenting on how existing models can be enriched with a small amount of labeled data from a different domain, to moderate content in that domain. Co-training with only 20% of the cost of having labelled data leads to an overall performance reduction of 2.9% . Therefore, co-training for domain adaptation can be considered a viable option when there is not enough labelled data.

Chapter 4

Data

4.1 DATASETS

4.1.1 Data Collection

- **Twitter1**
  Twitter is one of the most widely used platforms, almost rivalling Facebook therefore it has been the subject of numerous studies [26, 28, 34] on content moderation [1]. The dataset we use from Twitter for studying content moderation is from Gupta et al.'s study [28]. Posts were collected using hashtags related to protest, violence, and uprising. Both text and image from the posts are present. There are a total of 4671 posts.

- **Twitter2**
  There is another publicly available dataset [5] of Hate speech on Twitter. The csv files contain the Twitter ID of the posts. We rehydrated [2] the data by fetching the post content from Twitter API. Three classes of hate speech are present: racism, sexism and neither. Therefore racism and sexism posts are marked as positive and neither as negative.

- **Reddit**
  Reddit is a platform that contains multimodal data, so we searched for some dumps of a multimodal hate-related dataset. To the best of our knowledge, no such dataset is publicly available which contains the image and the caption related to posts for sensitive content. There are interest-centric communities on Reddit often called as subreddits. These subreddits contains post specific to a certain topic. Thus by identifying subreddits having

[1] Facebook has a highly restrictive API which makes data collection difficult therefore we do not use it in our study

[2] From Twitter ID posts are fetched from Twitter API

hate speech related content, we can extract the positive class for our problem. There is a list of quarantined subreddits, the subreddits which were banned to diminish hateful behavior on Reddit. There is an abundance of information in various previous works about text related hateful subreddits for eg r/fatpeoplehate, r/CoonToon etc [23] but none of them take into account pictorial aspects of a post.

We searched for a list of image related subreddits on [6]. We decided to work on the subreddit- r/pics. It is among the largest subreddit which contains images along with captions. Now to collect data from r/pics we used Reddit API. It gives only the top 1000 posts from a subreddit. We decided to further label the data on the basis of the number of upvotes, the number of downvotes, mean score to distinguish between positive and negative class. But due to the limited amount of data, models were not efficient. So, we collected data from a third-party API named as pushshift.io [7]. It gives N number of posts from a subreddit, where N is any positive integer. These third-party APIs does not give us the exact count of upvotes/downvotes as given by Reddit API. This is done to avoid spam as stated by [29].

After collecting a sufficient number of posts from pushshift, labels were assigned on the basis of score parameter. After running the classifiers, the performance was not as expected. So we filtered 100 posts and manually checked the labels. We found that the labeling on the basis of the score parameter was wrong. Hence, we considered r/pics as a positive class and started looking for some quarantined image subreddits data for the negative class. The authors of [30] decided to work on r/RoastMe. This is a subreddit where people humiliate others by adding pictures and comments. This subreddit got banned recently. We tried to collect image and text content from r/RoastMe, but as it got banned so data could not be fetched from the API. Similarly, we decide to look for data from quarantined subreddits like r/gore, r/watchpeopledie, etc which are multimodal subreddits containing sensitive content. But due to API restrictions, we were not able to collect data from quarantined subreddits. So, after going through some of the posts in multiple subreddits, we found subreddit- r/creepy contain some hateful content that suffices the need for positive class. Two annotators manually annotated 200 posts from r/creepy. The inter-annotator agreement was fair. Finally, we used subreddit r/creepy as a positive class and r/pics for the negative class.

- **Wikipedia1**

  Wikipedia is a well-known encyclopedia that allows members to modify the content. There is a Wikipedia talk page that allows editors to discuss the changes associated with the article. Editors use this sometimes for personal interests and converse in harsh language. There is a dataset made publicly available by authors of paper [45]. This dataset is collected from different Wikipedia talk pages. It contains 115K comments and labels corresponding to each comment. Label 1 represents a positive class and 0 represents a negative class. The dataset is available at [8].

- **Wikipedia2**

  There is another publicly available dataset of Wikipedia on Kaggle [9]. The toxic comment classification challenge aims to classify a given comment into 6 different classes of toxicity-Toxic, Severe Toxic, Obscene, Threat, Insult and Identity Hate. The dataset contains 159K comments. Each comment has six labels corresponding to each class. If a comment belongs to any one of the 6 categories, the corresponding class label is 1. As our main focus is on binary classification so if at least one of the labels corresponding to six classes is 1, we label that comment as positive else negative.

Dataset Name	Positive	Negative	Total	Positive class %	Text, image
Twitter1	3619	1052	4671	77%	Text + Image
Twitter2	1200	2000	3200	37%	Text
Reddit	2073	2598	4671	44%	Text + Image
Wikipedia1	647	5146	5793	11%	Text
Wikipedia2	783	7195	7978	10%	Text
Quora	817	12244	13061	6%	Text
Whisper	760	1720	2480	30%	Text

Table 4.1: Count of number of positive and negative samples for different datasets.

- **Quora**

  Quora is a well-known platform to ask questions and learn from each other. People can ask questions and other members can share their knowledge about the same. This helps them connect with others who contribute quality answers. A key challenge is to filter out insincere questions, those founded upon making false premises, or that intend to make a statement rather than look for helpful answers as stated by [10]. The question is categorized as insincere if it has a non-neutral tone or is disparaging or inflammatory or making false assumptions rather than stating the reality. The dataset [11] contains 1.3 million questions, each labeled as positive if it is insincere and negative otherwise.

- **Whisper**

  Whisper is an anonymous social media platform. This anonymity encourages to express thoughts without fear of recrimination. As stated by [43], there is a need to study how anonymity has changed the way of social interactions. There is a public dataset available at [12]. It is collected by [34] which contains 7K whispers all belonging to hate speech. This forms the positive class. Further, to collect data for negative class, we scrape data from whisper website [13].

Dataset	Text	Sensitive/Non-sensitive
Twitter1	#Dhaka's streets turn to rivers of blood as Muslims use car parks and garages to slaughter animals for Eid al-Adha	Positive
	nice to see that the top trending post by suriya #TamilNaduBandh #Saithan are located around TamilNadu	Negative
Twitter2	Deconstructed tart by lazy tarts #MKR	Positive
	Will someone pls assist Colin in the washing of his hair. Sorry Colin, big fan but pls... Some shampoo mate! #mkr	Negative
Reddit	i see your gibbons skeleton and raise you a pug skull!	Positive
	50 years of increasing obesity and diabetes. Thanks McDonalds, or should I say Ray Kroc.	Negative
Quora	I hate happy people laughing constantly on the bus. shut the fuck up!!!!! or I love God but I hate religious people.	Positive
	"I was camping in a park and the cops rolled up and arrested everyone but me because I gave a cop a bump of coke. In the end, it wasn't worth it."	Negative
Whisper	I hate black people. I know they are not all bad, but I'm sick of them blaming everyone for all their own ghetto ass problems.	Positive
	Being a single dad to a baby girl has made me jump out of my douche bag days. My days are now for making her feel beautiful and comfortable in her skin.	Negative
Wikipedia1	hey punk dont be deleting my stuff, you know nothing bout the harly drags so stay out of my shit you stupid nerd, punk fag female thats all u, bitch	Positive
	== Move of The Unbeatables == If you need a move reversed, but the redirect has a history, please note that the tag goes on the redirect, not on the article.	Negative
Wikipedia2	How can Black Lives Matter accomplish it's goals without looting and violent protesting?	Positive
	What is your opinion on the justification given by Zionists that the Jews had historically been belonged to Israel?	Negative

Table 4.2: Examples of a Positive and a Negative sample for each dataset.

4.1.2 Data Pre-processing

As the tweets from twitter dataset contain hashtags which specify the event names, so we anonymize these entities.
For example,
Tweet - "nice to see that the top trending post by suriya #TamilNaduBandh #Saithan are located around TamilNadu"
Anonymized Tweet - "nice to see that the top trending post by <NAME> are located around tamilnadu"

Text data
We converted complete text to lower case, removed hashtags and separators used to indicate different sentences line "newline", "nextline" etc. We used Named Entity Recognizer to replace names, places and dates by special tokens NAME, PLACE, and DATE respectively. Further, we filtered stop words for bag of words and TF-IDF models. For other models, we left the stop words as it is so that semantic meaning of a sentence is preserved.

Image data
As all the images were of different sizes, so to convert all of them to the same size we resized, normalized, cropped and augmented the images.

Chapter 5

Methodology

5.1 Baseline

Feature based Models:
We start with a bag of words vector and Term Frequency-Inverse Document Frequency(TF-IDF) weights of different words. Then top K features were used for classification. Support Vector Machine(SVM) and logistic regression(LR) classifier are used to classify posts.

Text Baseline:
As discussed in Chapter 2, we use [28] and [45] to test our model against. Also, we compare all the models for understanding the results on different datasets. We also used CNN based text classification models from [44]. These models are compared with LSTM, RNN and GRU deep learning models. [42] have used Capsule Networks, but only high level architecture details are mentioned, we cannot compare with their model.

Fusion Baseline:
For multi-modal classification, we extract features from a pre-trained VGG16 model and use fully connected layers on top. Also, a combination of LSTM and CNN from [28] is used to compare a different approach.

5.2 Feature Generation

For text data: We have used both TF-IDF and word2vec for feature generation from comments. Our vocabulary size was around 29K for twitter dataset, 180K for Wikipedia dataset, etc. and therefore the TF-IDF vectors for each comment is of very high dimension. As TF-IDF features do not take into account the contextual information, thus we also use word2vec embeddings.

Word2vec models are shallow two-layer neural networks that take the linguistic context of words into account to produce word embeddings. It takes as input a large corpus of text data and produces vector space of several hundreds of dimensions so that similar words are assigned vectors

in close proximity. There are many pre-trained word embeddings available that are trained on a large corpus. We have used glove embedding vectors.

For image data: The images extracted from Reddit and those from twitter dataset are of unequal dimensions. So a given image is cropped, augmented, resized and normalized to convert it into (224,224) dimension. PIL image library is used to perform the above four tasks. Further, if an image is not available for a post, a vector of (224,224) dimension containing zeros is assigned corresponding to that post id.

5.3 Text Model

LSTM (Long short term memory) architecture:
LSTM is composed of memory blocks called cells. Two states are transferred from one cell to another- hidden state and cell state. Cells are memory blocks as they remember the manipulation done with the help of three gates: input gate, output gate and forget gate. LSTM addresses the problem of vanishing gradients by using this three-gate structure.

Input gate helps to determine how much of the past information needs to be passed along to the future. Forget gate helps to determine how much of the past information to forget and The output gate reads the state from the memory and decides what the next state should be using a tanh activation function.

Capsule Network Architecture:
Capsule encapsulates all important information about the state of the feature they are detecting in vector form. Different capsules are stacked one after another. Information is passed from one capsule to another using dynamic routing protocol. Lower level capsule will send its input to the higher level capsule that "agrees" with its input. This is done using dynamic routing algorithm [37]. This routing is performed between two Capsule.

Convolutional Neural Networks(CNN) used for text do not capture the hierarchial patterns in text data. For example, if two documents have exactly similar sentences but in a random order, CNN's won't learn this. Also pooling layers do not capture semantic similarity. Further, representing words in one-hot encoding form can lead to substantial data sparsity and hence, poor model performance. Thus capsule networks help solve this problem using dynamic routing.

The different layers in its architecture are as follows:
Input Layer -
The input post is tokenized into words and each word is converted to word embeddings. These embedding vectors are passed as input to next layer.

Embedding Layer -
In the embedding layer, the words are converted to lower-dimensional vectors by converting them to word embeddings. The main advantage of using word embeddings is that they are able to capture context similarity and due to their smaller dimensionality, they are fast and efficient for deep learning and NLP tasks. The word embeddings of each word in the sentence are concatenated and fed to the hidden layer.

Hidden Layer -
This layer extract n-gram features at different positions of the sentence using LSTM. It extracts features from input vectors.

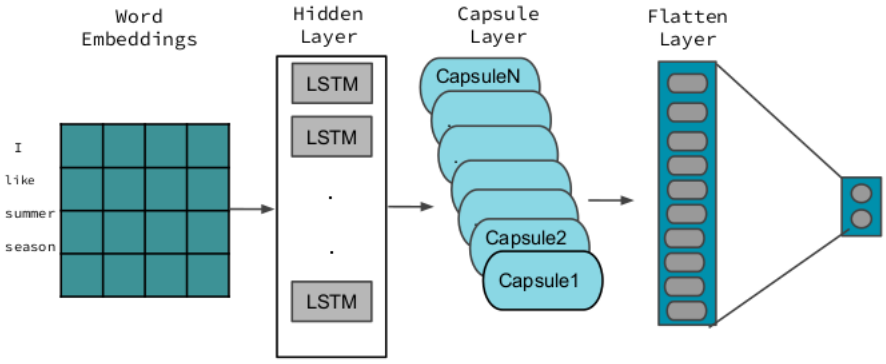

Figure 5.1: Capsule Network Architecture. The text data is processed and is converted into embeddings. LSTM extracts features from embedding vectors. Capsule learns the representations. Final layer apply softmax and produces output.

Capsule Layers -
It contains n number of capsules. These capsules have the ability to learn the semantic meaning of the text by representing values as vector and not scalars. Only relevant features identified are learned by Capsules. This is done by dynamic routing [37].

Fully Connected Layer -
This is the final layer of the network which flattens and combines the high-level features that are learned by the previous layers. This layer also consists of the dropout layer for regularization. The output of this layer is fed to the softmax output layer for prediction.

5.4 Co-training

In practical scenarios, a large amount of unlabeled data is available. Manual annotation of this data is a cumbersome task. So to label such data, co-training algorithm can be used. Co-training is a machine learning algorithm, which is used when there is small amount of labelled data and

large amounts of unlabelled data. The algorithm [22] start with a set of labeled data and train a model on this data. Now, labels are predicted for the unlabeled data by the classifier. If the unlabelled data is of the same domain as that of labelled data then the predictions are prone to less error. However, if the domain of labelled and unlabelled data is different the model does not generalize well on unlabelled data. For example, if labelled data is of Twitter and unlabelled data is of Reddit. As both platforms have variations in the vocabulary being used, size of the post, etc. so the model trained on Twitter does not adapt well on Reddit. To bridge this gap "Co-training for Domain Adaptation" [24] can be used where model can adapt to different domains data.

In our case, we have data from 5 different domains corresponding to each platform. As we are aiming to provide solution to the problem of content moderation, where we need to provide cost-effective solution which can work for data from all domains.

In co-training for domain adaptation algorithm, the data on which model is trained is called source data and the one for which domain adaptation needs to be done is called target data. We start by training model on source data. With each iteration we add some x percent(in our case x is 20) of target data to the training data, lets denote this as intermediate data. Next we shuffle the intermediate data and again train the model. This process is repeated for a few iterations. After this labels are predicted for rest(excluding samples added to source data) target data.

Algorithm 1 Co-training for Domain Adaptation

Result: Evaluation metric values for Y_U

Given:

$\boldsymbol{X}$ = $Domain_1$ data samples

$\boldsymbol{Y}$ = $Domain_2$ data = $Y_L \cap Y_U$

where Y_L is labelled data used to augment existing $Domain_1$ model and Y_U is unlabelled data used for testing

iterations = K

model(X)

Let r = n(Y_L) / K

$\boldsymbol{Z}$ = r of Y_L

Y_L = Y_L - Z

for *each* ***i*** *in* ***iterations*** **do**

  intermediate_data = X + Z

  Train model(intermediate_data)

  Test on Y_U

  X = X + Z

  Z = r of Y_L

  Y_L =Y_L - Z

end

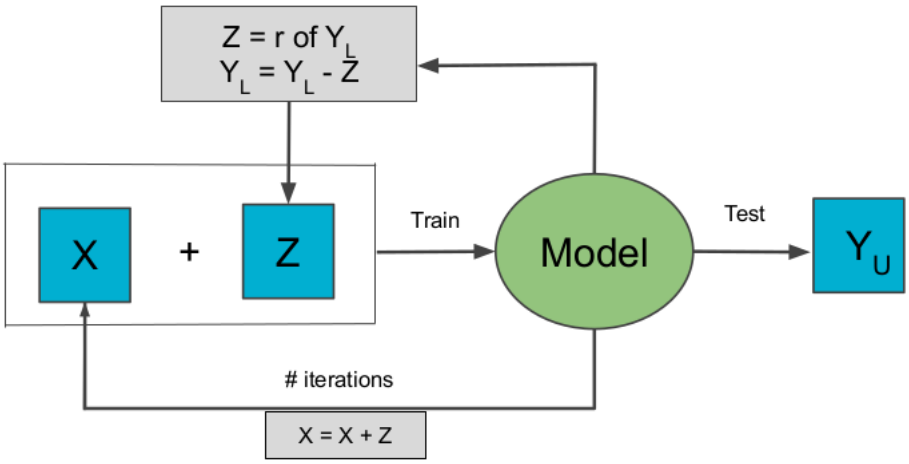

Figure 5.2: Co-training for Domain Adaptation diagram.

In Figure 5.2 initially X denotes all data samples from Domain1 Y_L denotes labelled data samples from Domain2 and Y_U denotes unlabelled data samples from Domain2. If we add 20% data and there are 5 iterations(iterations=5), then r = 20/5 i.e. r=4%.
In each iteration to X we are adding r samples from Y and training the model.
Predict labels for Y_U.
X increases in size after each iteration.
After iterations this process is stopped, final results on Y_U is calculated by averaging the results over each iteration.

5.5 Fusion Models

The model is reused form [28]. They have used LSTM for text classification and a combination of Object recognition and Scene recognition models followed by Global Average Pooling(GAP) layers and fully connected layers for image classification. Fixed sized vector of text and image models is combined. This is fed to Fully connected layer with sigmoid activation.

In new fusion model, we replaced LSTM text model with CapsNet model, and reused the rest of the model as discussed above.

Chapter 6

Experiments and Results

Chapter 5 dealt with different methods used for experimenting various datasets. Now, this chapter provides a comparative study of different approaches used.

6.1 Methodology

We report 5 fold cross-validated results on all methods. The positive class for all models is the one containing sensitive content and negative class containing non-sensitive content.

6.2 Experimental Metrics

Considering class imbalance problem with various datasets, we used Macro-F1 score, Accuracy and area under Precision-Recall curve(PR-curve) measures for classification.

PR curve also known as average precision, measures the precision against recall. It tells us how good the classifier is in predicting positive class in case of class imbalanced datasets. Thus we use this measure to report our models performance.

Macro-F1 is a variation of F1-score, which is reported to evaluate classification results for skewed datasets. Macro-F1 values are calculated by first computing F1-scores for each class in isolation and then averaging over all the classes.

$$\text{F1(pos)} = \frac{\text{2*P(pos)*R(pos)}}{\text{P(pos) + R(pos)}} \tag{6.1}$$

$$\text{F1(neg)} = \frac{\text{2*P(neg)*R(neg)}}{\text{P(neg) + R(neg)}} \tag{6.2}$$

$$\text{F1 macro} = \frac{\text{F1(pos) + F1(neg)}}{2} \tag{6.3}$$

		Predicted	
		Positive	Negative
Actual	Positive	a	b
	Negative	c	d

From above table, we can compute P(pos) and R(pos) as follows:

$$P(pos) = \frac{a}{a+c} \tag{6.4}$$

$$R(pos) = \frac{a}{a+b} \tag{6.5}$$

Similarly, we can compute P(neg) and R(neg).

6.3 Supervised Models Results

6.3.1 Text model

Seven different methods have been used namely Capsule Network referred as CapsNet, Gated Recurrent Units referred as GRU, Logistic Regression referred as LR_machina as the code is available so we use it to compare with our method, LR_Badjatiya, as code is available so we compare our methods with their methods, Long Short Term Memory referred as LSTM, Multi Layer Perceptron referred as MLP, Random Forest Referred as RF and Support Vector Machines referred as SVM.

For all the above methods results on Twitter2 dataset are reported in Table 6.1.[1] For simplicity, we choose to discuss results only on one dataset as it help us to illustrate main findings from our analysis. We have added other tables in Appendix. As can be inferred from the table, CapsNet performs better than all the other methods.

[1]We found similar results for all other datasets.

Dataset	Method	Accuracy	Macro F1	Area ROC curve	Area PR curve
Twitter2	CapsNet	0.8203	0.8254	0.8251	0.7695
	CNN	0.7960	0.6400	0.7931	0.6960
	GRU	0.7168	0.6977	0.6935	0.6983
	LR_Badjatiya	0.7970	0.7560	0.7523	0.7260
	LR_machina	0.7281	0.6772	0.7346	0.3805
	LSTM	0.7303	0.7076	0.7033	0.7057
	MLP	0.7468	0.7300	0.5593	0.6916
	RF	0.6781	0.5536	0.7454	0.1784
	SVM	0.7531	0.7314	0.7401	0.5309

Table 6.1: Results of all text models on Twitter2 dataset. CapsNet produces highest Area under PR curve than all other methods.

Our CapsNet method outperforms LSTM method results by 10% in average precision for all datasets from Table 6.2. We note that same model can be interpreted very differently depending on the method. Also, for datasets from similar domain, Wikipedia1 and Wikipedia2 we can see similar results. Area PR curve is taken as a measure to report the results.

Dataset	Method	Accuracy	Macro F1	Area ROC curve	Area PR curve
Quora	CapsNet	0.9494	0.6959	0.9496	0.9269
	LSTM	0.933	0.6731	0.4095	0.656
Reddit	CapsNet	0.7952	0.7967	0.7967	0.7373
	LSTM	0.733	0.7306	0.6217	0.7321
Twitter1	CapsNet	0.8190	0.7953	0.8365	0.7695
	LSTM	0.7854	0.8635	0.9117	0.6748
Twitter2	CapsNet	0.8203	0.8254	0.8251	0.7695
	LSTM	0.7303	0.7076	0.7033	0.7057
Whisper	CapsNet	0.9798	0.9856	0.9778	0.9783
	LSTM	0.9824	0.9816	0.9654	0.9816
Wikipedia1	CapsNet	0.943	0.8361	0.9439	0.9195
	LSTM	0.924	0.7775	0.4274	0.7413
Wikipedia2	CapsNet	0.943	0.8361	0.9439	0.9195
	LSTM	0.9414	0.8098	0.6943	0.7698
Total Average	**CapsNet**	**0.8928**	**0.8244**	**0.8962**	**0.8600**
	LSTM	0.8613	0.7919	0.6761	0.7516

Table 6.2: Text classification results on 7 datasets. Overall CapsNet perfromed better than LSTM.

6.3.2 Fusion Model

Our fusion model combines CapsNet model for text with object and scene model for image, it is referred as CapsFusion. We compare this model against baseline model(LstmFusion). It combines LSTM model for text with object and scene model for image.

Dataset	Method	Accuracy	Macro F1	Area ROC curve	Area PR curve
Twitter1	LstmFusion	0.8319	0.6711	0.6626	0.8381
	CapsFusion	**0.8458**	**0.6968**	**0.6793**	**0.8613**
Reddit	LstmFusion	0.7548	0.7529	0.7831	0.7566
	CapsFusion	**0.8169**	**0.8141**	**0.7128**	**0.8149**

Table 6.3: Fusion model results on Twitter1 and Reddit datasets. CapsFusion gives better results than LstmFusion by 5% in average precision.

6.3.3 Co-training results

As discussed in Section 5.4, The source dataset is Twitter1 and target datasets are Twitter2, Reddit, Wikipedia1, Wikipedia2, and Quora and Whisper. Taking complete source data as input and with each iteration of cross-validation, we add 5% samples of target data. We shuffle this data and then train the model. The labels for target data is predicted. Co-training results are evaluated on LSTM and CapsNet as they performed better than all other models for text classification.

After obtaining results on addition of 5% domain data, we observed that Reddit, Twitter2 and whisper performed better as the total number of samples in the original dataset is less. For Wikipedia1, Wikipedia2 there is a huge difference between the number of samples of positive and negative classes. So we randomly sample 5% of data in equal ration from both the classes. Quora dataset contains 13K samples, so we tested the co-training results by adding 20% of the dataset. We found that the results are better than those obtained by adding 5% data.

Therefore, we analyze the accuracy of different co-trained models by gradually increasing the target domain dataset samples. For multimodal dataset, experiments are performed in two ways:

- By taking Twitter1 as source dataset and Reddit as target dataset.
- By taking Reddit as source dataset and Twitter as target dataset.

We considered 20% of target domain data for each case. LstmFusion and CapsFusion models are trained.

From Table 6.5, we can infer that CapsFusion model performed better than LstmFusion model.

Figure 6.6 shows the comparison of top two models trained on complete labelled data with co-trained models trained on 20% data.

Trained on	Co-trained on	Method	Accuracy	Macro F1	Area ROC curve	Area PR curve
Twitter1	Twitter2	LSTM	0.6395	0.6105	0.6052	0.6091
	Reddit		0.6679	0.6673	0.7211	0.6748
	Wikipedia2		0.8944	0.6951	0.4785	0.6881
	Wikipedia1		0.7120	0.6089	0.5967	0.5956
	Quora		0.8758	0.5917	0.2889	0.6202
	Whisper		0.8420	0.8743	0.8311	0.8802
Twitter1	Twitter2	CapsNet	0.6489	0.6321	0.6586	0.6232
	Reddit		0.6691	0.669	0.68	0.5581
	Wikipedia2		0.8988	0.7341	0.3801	0.7457
	Wikipedia1		0.8983	0.7496	0.4269	0.748
	Quora		0.9236	0.6739	0.2399	0.6723
	Whisper		0.8582	0.9167	0.8921	0.9201

Table 6.4: Co-training on text datasets. Cotrained models are able to learn data from different domain well.

Trained on	Co-trained on	Method	Accuracy	Macro F1	Area ROC curve	Area PR curve
Twitter	Reddit	LstmFusion	0.6715	0.6737	0.7201	0.6004
		CapsFusion	**0.7174**	**0.7155**	**0.7173**	**0.6049**
Reddit	Twitter	LstmFusion	0.5394	0.5763	0.7944	0.5103
		CapsFusion	**0.7444**	**0.623**	**0.9045**	**0.6263**

Table 6.5: Co-training on image datasets. Cotrained image models learn different domain images well.

6.4 Qualitative analysis

Error Analysis

Taking 50 data points which are classified correctly by our model, and classified incorrectly by other models, we tried to manually analyse the false positives and false negatives. False positives are the example points which have original label as negative but predicted label as positive.

Dataset	Method	Accuracy	Macro F1	Area ROC curve	Area PR curve
Quora	CapsNet	0.9494	0.6959	0.9496	0.9269
	Cotrain_CapsNet	0.9236	0.6739	0.2399	0.6723
	LSTM	0.933	0.6731	0.4095	0.656
Reddit	CapsNet	0.7952	0.7967	0.7967	0.7373
	Cotrain_CapsNet	0.6691	0.669	0.68	0.5581
	LSTM	0.733	0.7306	0.6217	0.7321
Twitter1	CapsNet	0.8190	0.7953	0.8365	0.7695
	Cotrain_CapsNet	-	-	-	-
	LSTM	0.7854	0.8635	0.9117	0.6748
Twitter2	CapsNet	0.8203	0.8254	0.8251	0.7695
	Cotrain_CapsNet	0.6489	0.6321	0.6586	0.6232
	LSTM	0.7303	0.7076	0.7033	0.7057
Whisper	CapsNet	0.9798	0.9856	0.9778	0.9783
	Cotrain_CapsNet	0.8582	0.9167	0.8921	0.9201
	LSTM	0.9824	0.9816	0.9654	0.9816
Wikipedia1	CapsNet	0.943	0.8361	0.9439	0.9195
	Cotrain_CapsNet	0.8983	0.7496	0.4269	0.748
	LSTM	0.924	0.7775	0.4274	0.7413
Wikipedia2	CapsNet	0.943	0.8361	0.9439	0.9195
	Cotrain_CapsNet	0.8988	0.7341	0.3801	0.7457
	LSTM	0.9414	0.8098	0.6943	0.7698
Total Average	**CapsNet**	**0.8928**	**0.8244**	**0.8962**	**0.8600**
	Cotrain_CapsNet	0.8161	0.7292	0.5462	0.7114
	LSTM	0.8613	0.7919	0.6761	0.7516

Table 6.6: Comparison of Cotrained models with models trained on complete labelled data.

Similarly, false positives are the example points which have original label as positive but predicted label as negative.

For text - CapsNet and LSTM models are taken. Training the CapsNet and LstmFusion model on Twitter1 dataset, we predicted the labels for 500 samples. Then we compared original and predicted labels. If original and predicted labels are similar, the samples are discarded. From rest of the samples, we take 50 false positive samples and 50 false negative samples for analysis.

For image - CapsNet fusion and LSTM fusion models are trained on Twitter1 dataset. Labels are predicted for 300 samples. After separating 25 samples each for false positives and false negatives, we manually annotate the samples.

DeepSHAP (Deep Learning SHapely Additive exPlanations) [14] is used to explain the output of any deep learning model. We analyze two examples one for positive and another for negative. The results are as shown below in Figure 6.1 and 6.2.

Text - Where are the activists and foot soldiers when k'tak bleeds in silence.
Label - Positive

Figure 6.1: DeepSHAP results on Positive content. The words foot, bleeds, silence are related to sensitive content.

Text - The proud hero of kashmir! The hero of freedom struggle.
Label - Negative

Figure 6.2: DeepSHAP results on Negative content. The words proud, freedom, hero are related to non-sensitive content.

The words shown in red contribute more towards models output than blue words. The value above the given word shows the contribution of the word.

From Figure 6.1 the words shown by red i.e. bleeds, silence, foot are contributing more towards the output as should be the case, as the text content is sensitive. Bleeds contribute the highest, then silence and then foot. Similarly, in Figure 6.2 the words shown by red i.e. proud, hero, freedom are contributing more towards the output as should be the case, as the text content is non-sensitive. Proud contributes highest, then hero and then freedom.

6.5 Co-training Tradeoff Analysis

As we gradually increase the percentage of points which are added during co-training, how the model accuracy varies. To analyze this behaviour, we trained LR, CapsNet and LSTM model for text datasets. LstmFusion and CapsFusion model for multimodal datasets.

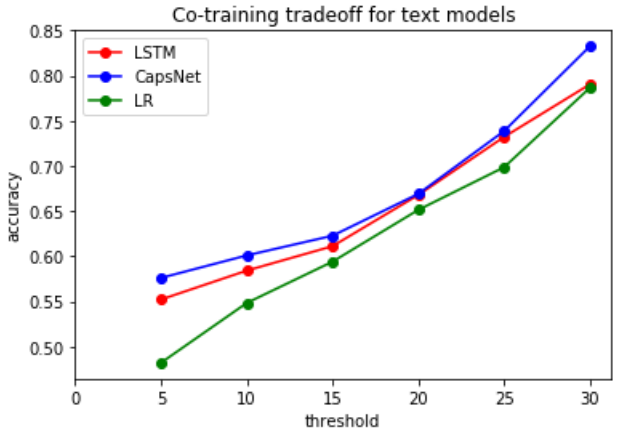

Figure 6.3: Co-training threshold analysis for text. CapsNet have high accuracy for all threshold values.

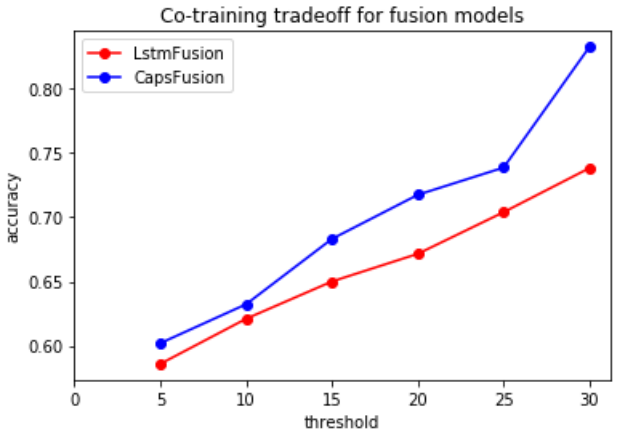

Figure 6.4: Co-training threshold analysis for multimodal data. CapsFusion have high accuracy for all threshold values.

Summary of results :

- **Class imbalance affect the performance of different methods**
  As we can see from table 4.1, the percentage of positive class samples in all the datasets is variable. Also different datasets have different number of total samples.

- **Average length of posts on different social media platform varies**
  The average post length on Twitter is 30, for Reddit is 10, for Wikipedia it is 50 etc. So it is hard to come up with some set parameters which will work for all platforms.

- **The style or way by which posts are expressed on different platforms also affect the classification**
  As Twitter has sparse language text while Reddit has well-formed communication patterns that resemble web blogs, so a similar text processing approach cannot be followed over all datasets.

- **Same social media text can be interpreted very differently depending on the choice of a method used for analysis**
  As in our case, bag-of-words is used for Logistic Regression and word embeddings is used for LSTM and Capsule Networks.

- **Computational requirements**
  Concerning computational requirements, it is fair to say that CapsNet is significantly slower than the other algorithms, as each iteration is of high complexity.

Chapter 7

Portal

We design this portal as a proof-of-concept for different methods.
Technologies used: Python, Keras, Tensorflow, sklearn, Flask, Tweepy, Twitter API, Pandas, Matplotlib

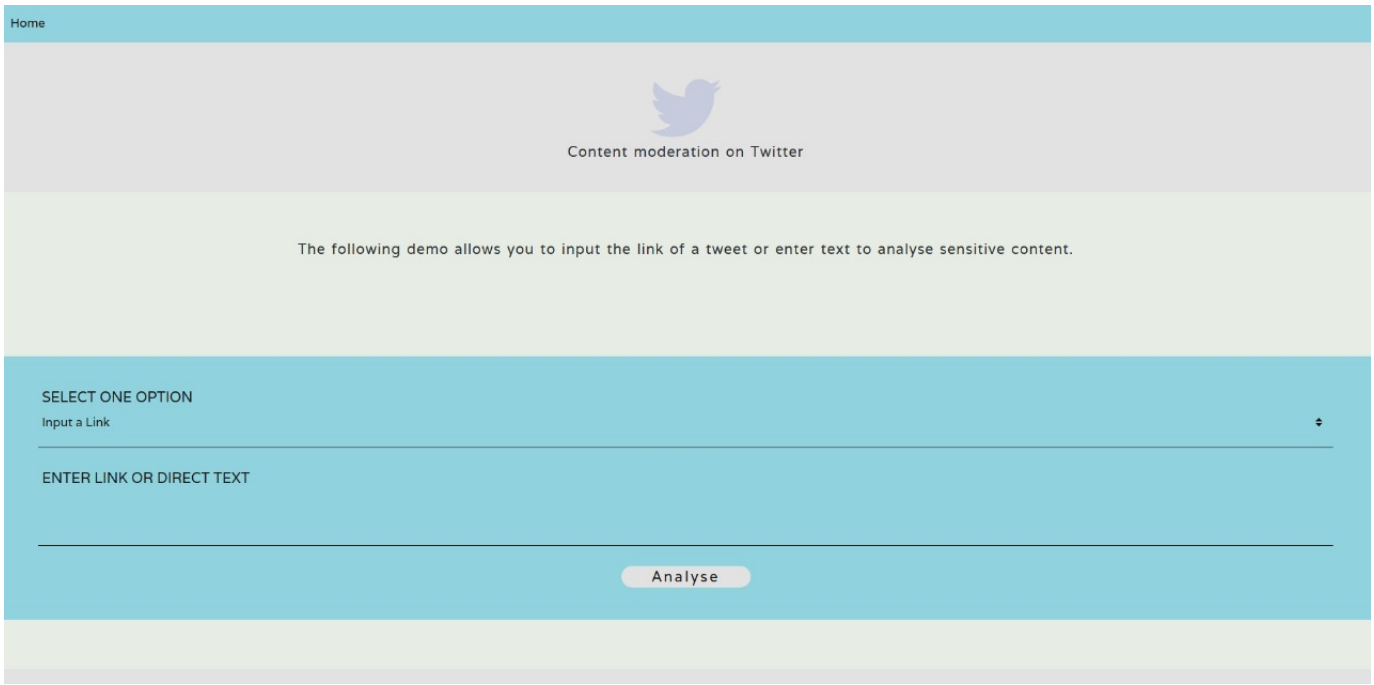

Figure 7.1: Portal to analyze sensitive content.

The select one option allows the user to choose :

- Input a link
- Input text

The app allows to user to analyze a tweet containing only text or both text and image through the "Input a link" option. The user can also choose to analyze pure text through the "input text" option. This app analyzes the text and the image and returns "Positive" or "Negative", depending upon the content.

The flow of the App

When the user enters the link to the tweet, The text and the image(if present) is extracted from the link and saved using twitter API. Depending on the type of data (text/image) the models run in the backend and returns the result to the web page (front-end).

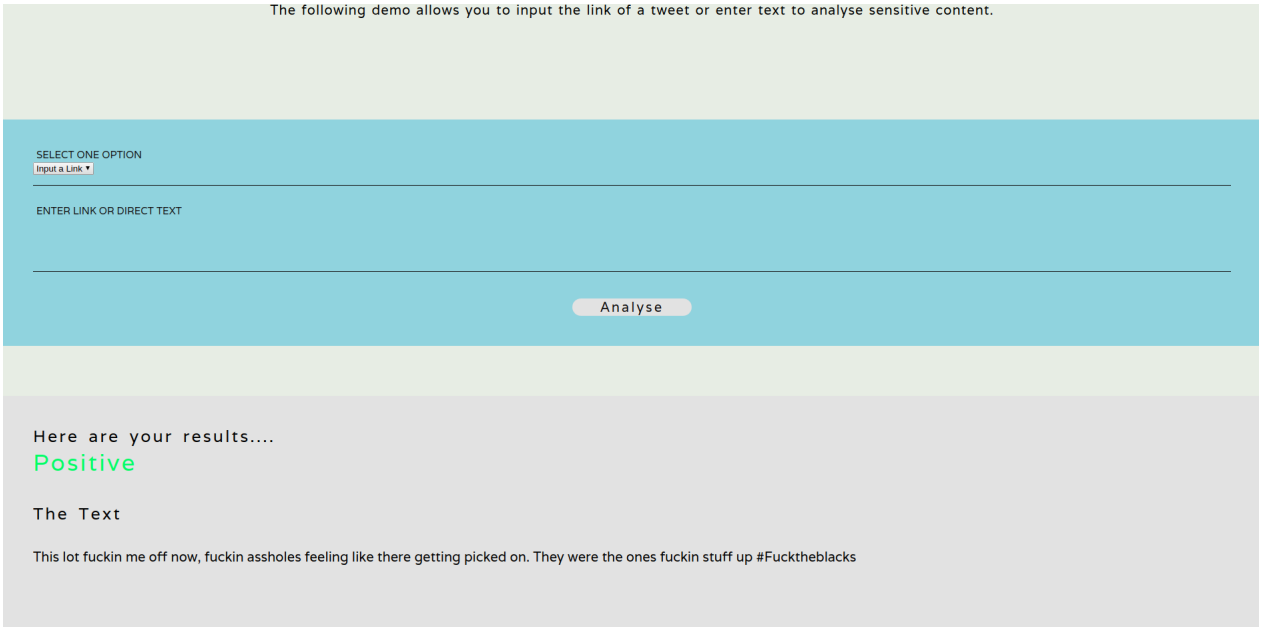

Figure 7.2: Example of a post.

Chapter 8

Conclusions, Limitations, Future Work

8.1 Conclusions

In this study, we present a classification scheme for content moderation across different platforms. CapsNet model is proposed which in a novel idea. For this, we first started by analyzing results on text models. After the experimental analysis we found that LSTM and CapsNet worked fairly well on different datasets. We observed an increase of 10% in average precision using our approach.

We benchmark different methods for content moderation on different platforms. Co-trained models performed well, and they can be used as a cost-effective solution for annotating unlabelled data. Our preliminary investigations on how co-training can be used to improve domain adaptability shows good results.

8.2 Limitation

Considering the computational requirements, it is fair to say that CapsNet is significantly slow than the other algorithms. Not taken into account user features. Limited computations so we reduced the dataset on which models are trained. We faced difficulty is in getting hold of real world data for searching for subredddits containing multimodal data. On image related platforms like Pinterest and Instagram, it is difficult to get data and further difficult to get the data annotated.

8.3 Future Work

As a part of future work, we can explore co-training on performed. This can be done by taking each dataset as source and co-training it on rest of the datasets. Currently binary classification,

can be extended further to multi-class classification. No feature engineering on different datasets as we want to see results on all datasets. Providing better explanation through shap and why certain post is marked worthy of content moderation can be examined further.

Bibliography

[1] https://www.wired.com/story/the-punishing-ecstasy-of-being-a-reddit-moderator/.

[2] https://medium.com/nanonets/content-moderation-in-2019-human-vs-ai-1c7993e5e4f3.

[3] https://medium.com/acm-cscw/you-cant-stay-here-the-efficacy-of-reddit-s-2015-ban-examined-through-hate-speech-93f22b140f26.

[4] https://www.theverge.com/2015/5/14/8606923/reddit-anti-harassment-policy.

[5] https://github.com/zeerakw/hatespeech.

[6] https://www.reddit.com/r/ListOfSubreddits/wiki/listofsubreddits.

[7] https://pushshift.io/.

[8] https://figshare.com/articles/Wikipedia_Detox_Data/4054689.

[9] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data.

[10] https://www.kaggle.com/c/quora-insincere-questions-classification/overview.

[11] https://www.kaggle.com/c/quora-insincere-questions-classification/data.

[12] https://github.com/Mainack/hatespeech-data-HT-2017.

[13] http://whisper.sh/.

[14] https://github.com/slundberg/shap.

[15] 8 out of 10 indians have faced online harassment. https://www.thehindu.com/news/national/8-out-of-10-indians-have-faced-online-harassment/article19798215.ece, 2013.

[16] Facebook the trauma floor. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona, 2013.

[17] How much time spent on social media. http://www.bbc.com/future/story/20180118-how-much-is-too-much-time-on-social-media, 2013.

[18] The positives of social media: Spread of information. http://lifeasoflate.com/2013/11/the-positives-of-social-media-spread-of-information.html, 2013.

[19] We are numbed to the harm of digital insults. https://www.technologynetworks.com/neuroscience/news/we-are-numbed-to-the-harm-of-digital-insults-311996, 2013.

[20] Wikipedia no personal attacks. https://en.wikipedia.org/wiki/Wikipedia:No_personal_attacks, 2013.

[21] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *WWW*, 2017.

[22] Avrim Blum and Tom Michael Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, 1998.

[23] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *PACMHCI*, 1:31:1–31:22, 2017.

[24] Minmin Chen, Kilian Q. Weinberger, and John Blitzer. Co-training for domain adaptation. In *NIPS*, 2011.

[25] Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *ICWSM*, 2017.

[26] Luis Gerardo Mojica de la Vega. Determining trolling in textual comments. 2017.

[27] Spiros V. Georgakopoulos, Sotiris K. Tasoulis, Aristidis G. Vrahatis, and Vassilis P. Plagianakos. Convolutional neural networks for toxic comment classification. In *SETN*, 2018.

[28] Divam Gupta, Indira Sen, Niharika Sachdeva, Ponnurangam Kumaraguru, and Arun Balaji Buduru. Empowering first responders through automated multimodal content moderation. *2018 IEEE International Conference on Cognitive Computing (ICCC)*, pages 1–8, 2018.

[29] Jack Hessel, Lillian Lee, and David M. Mimno. Cats and captions vs. creators and the clock: Comparing multimodal content to context in predicting relative popularity. In *WWW*, 2017.

[30] Anna Kasunic and Geoff Kaufman. "at least the pizzas you make are hot": Norms, values, and abrasive humor on the subreddit r/roastme. In *ICWSM*, 2018.

[31] Srijan Kumar, William L. Hamilton, Jure Leskovec, and Daniel Jurafsky. Community interaction and conflict on the web. In *WWW*, 2018.

[32] Jesse Mawhinney. 37 visual content marketing statistics you should know in 2016. http://blog.hubspot.com/marketing/visual-content-marketing-strategy, 2016.

[33] Shruthi Mohan, Apala Guha, Michael Harris, Fred Popowich, Ashley Schuster, and Chris Priebe. The impact of toxic language on the health of reddit communities. In *Canadian Conference on Artificial Intelligence*, pages 51–56. Springer, 2017.

[34] Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. A measurement study of hate speech in social media. In *HT*, 2017.

[35] Edward Newell, David Jurgens, Haji Mohammad Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths. User migration in online social networks: A case study on reddit during a period of community unrest. In *ICWSM*, 2016.

[36] Hasso Plattner. Aggression identification using deep learning and data augmentation. 2018.

[37] Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. In *NIPS*, 2017.

[38] Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard J Jansen. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[39] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *SocialNLP@EACL*, 2017.

[40] Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. Analyzing the targets of hate in online social media. In *ICWSM*, 2016.

[41] Shubham Singh, Rishabh Kaushal, Arun Balaji Budur, and Ponnurangam Kumaraguru. Kidsguard: Fine grained approach for child unsafe videorepresentation and detection. In *SAC*, 2019.

[42] Saurabh Srivastava, Prerna Khurana, and Vartika Tewari. Identifying aggression and toxicity in comments using capsule network. In *TRAC@COLING 2018*, 2018.

[43] Gang Wang, Bolun Wang, Tianyi Wang, Ana Nika, Haitao Zheng, and Ben Y. Zhao. Whispers in the dark: analysis of an anonymous social network. In *Internet Measurement Conference*, 2014.

[44] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *SRW@HLT-NAACL*, 2016.

[45] Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Ex machina: Personal attacks seen at scale. In *WWW*, 2017.