Towards Decentralized Moderation of Community Media Content

Summer Undergraduate Research Award

Ayush Patel
2016CS10396
Computer Science
CGPA: 9.305
Mob: 9891052662
cs1160396@iitd.ac.in

Mohit Gupta
2016CS50433
Computer Science
CGPA: 9.579
Mob: 9466479674
cs5160433@iitd.ac.in

Supervisor:-
Aaditeshwar Seth
Professor
Department of CSE
aseth@cse.iitd.ac.in
IIT Delhi

Prof. S. Arun Kumar
Head of Department
Department of CSE
sak@cse.iitd.ernet.in

1 Introduction

History proves that right from the starting of the time, Media had played an important role in influencing various sort of social, economical and political phenomenons in societies. It has led to union and division of society into various groups based on their opinions on different topics. It has led to many revolutionary movements through the masses. But some important question which arises are "Are all these news true?", "Can media even influence us to act opposite to our opinion?". Everyday we come across various fake news in various unmoderated social media like Facebook and Twitter which leads to formation of various echo chambers. These tend to make false images of people and groups among the society. But, moderated social media which tries to ensure that the information is credible and that it is presented with diverse viewpoints being covered, is hard to scale. This project aims to decentralize the moderation process so that it is easier to ensure the correctness and unbiasedness of the news been delivered.

This model consists of a community based news platform where the news is collected through different user groups who record their audio over IVR(Interactive Voice Response) systems. Before this audio is published, it is made sure that it is free of various issues viz-a-viz it is not abusive, it is not politically motivated etc. This working is made efficient using two levels of decentralization of moderation namely - community representatives (volunteers) and content moderators.

Community representatives forms the first level of decentralization which is closest to the user group. They listen to the audios recorded over IVR and then pass the relevant content to the moderator. Then the content moderators, who form the second layer of decentralization, check the content for any political and social biasedness which is procedure for the selection of content, and then they rank the content based on it's importance and relevance. Then the overall ranking of content is determined by a ranking algorithm.

It is also necessary to ensure that the various levels of moderation are working efficiently. Our ground truth for the correctness of the decision taken by volunteers is based on the decision taken by the moderators for the same content. Based on this, the volunteers are provided scores based on a scoring algorithm. These scores play an important role in determining the incentives of the volunteers. So, to get better incentives, the volunteers will try to increase their score and in turn work properly. This will increase the efficiency of our system.

A system can be said to be successful only if it suffices for the need of the users. To measure the popularity of the content among the users, a mathematical measure called Popularity Index is determined based on the lifetime of a content viz-a-viz whether people are listening to the content or skipping it etc. Then while publishing, the content which has high popularity index is preferred over the one with less popularity index. This system

provides moderated unbiased content to the user based on their preference.

We also aim to add certain features to our app by which the stories contributed by the community reporter are better organized and hence aid him to follow-up on the issue much more effectively. Data and information in such an organized form can then also be shown to various concerned authorities, hence making the voices of the people to be heard where they need to be and improve chances of a resulting impact. These tags help grouping similar kind of data and thus, a particular moderator specialized in that topic can go for the second level of screening of the same.

This project focuses on **Building election algorithms to solve the two problems of *content selection* and *content ranking*. From an optimization point of view, how many ratings per content should we get? How can we calculate the reputation of the volunteers in giving good scores?**

2 Objectives

Our main objective in this project is to design, build and validate an app to moderate and curate the voice messages recorded on IVR (Interactive Voice Response) systems incorporated in Mobile-Based Community Platforms.

The main objectives of the app is to to decentralize four decision points:

- **Selection of content:** The app focuses on the "editorial problem" of selection of content to make sure that it is not abusive, it is not politically motivated, etc and the information is complete. Before the audio is published at any Mobile-based community platform, we aim to make it sure that it is free of various issues viz-a-viz it is not abusive, it is not politically motivated, abusive, motivated by rumours etc. We need to improvise a method to handle the situations when multiple stakeholders participate and provide views on a given topic, and make the message becomes more complete in terms of its coverage of different aspects.

- **Ranking of content:** Through this model, the content moderators can easily rank the content based on it's importance and relevance. The overall ranking of content is then determined by a ranking algorithm.

- **Content Popularity Index:** An important task in any media based platform is to measure the popularity of their content among the users and simultaneously incorporate necessary changes to increase the same. The app will also provide indicators to the volunteers of the responses of the content by users i.e. after it was published, are people listening to it or skipping it. If a lot of people are skipping it then the volunteers should look more deeply into it, or the moderators can be consulted to investigate, and take corrective steps.

- **Organizing the bulk of data:** We also aim to add certain features to our app by which the stories contributed by the community reporter are better organized and hence aid him to follow-up on the issue much more effectively.

The app aims to satisfy the need of automatic content moderation in various Community Platforms by adding features like abusive content filters, Popularity Index, associates tags based on the broad topic/theme(and subtopics having to rank) the audio belongs to and thus, to organize audios collected via the app into issues and enable better tracking of the status of each.

3 Related Work

- Some social media like **Facebook** have worked on a method of content moderation known as **User Generated Censorship**. This method basically involves the users setting the flag or giving up/down votes to a post based on which an army of moderators block the content if found inappropriate. Due to lack of human force, the army is also supplemented by AI. This method has failed many times when someone uses a script/bot to automate the down-voting of a post in order to suppress some specific content.

  Eg. Suppressing of J30Strike news on Facebook.

- **Slashdot**, a news and commentary site dedicated to technology issues, uses **Distributed Moderation System** which is an extension of *User Generated Censorship* in which the weightage of the votes of users depend on their points which in turn depends on their previous comment/post history. Apart from this, it also has paid moderators who have unlimited moderation points. This turned out to be much better than *User Generated Censorship*.

- **YouTube** has been dabbling with a similar community-based moderation system. Called **"YouTube Heroes,"** it empowers an elite group of users to police the site and help weed out the flood of horrific videos that you can imagine gets uploaded every second to the popular platform

- **YouNow**, a social live-streaming platform that's primarily targeted at teens, has a fairly robust content moderation system. The site also employs proprietary technology to filter out such content–both videos flagged by users as well as those detected via its algorithmic system. It does that partly by analyzing comments in real time, which helps detect content that should be flagged.

- **Instagram** too is known for its heavy-handed content moderation, via such methods as **blocking the hashtags** for groups and communities that share potentially harmful content. Through this method, it has successfully blocked content that promotes eating disorders. How Instagram decides which hashtags to block remains a mystery.

This is something many social media researchers have looked into, but given that hashtags are a prime way for users to connect, these blocks help reduce the flood of potentially abusive content.

4 Approach to the project

The overall approach is to provide the content recorded by the user group over IVR to the community representatives (volunteers) who pass the relevant audios to the content moderators. These content moderators do the selection and ranking of the content based on various factors such as biasedness of the content, it's political impact etc. The overall ranking of the content is determined with the help on a ranking algorithm. Then the content finally been published is available to the user on demand. Then based on the lifetime of the content, a mathematical measure called Popularity Index of content is determined through which the community representatives get to know the need of user group and then refine the content accordingly. To ensure the efficient working of the volunteers, they are given scores based on their work which in turn determines the incentive that they get for their work.

- **Collection of Content**

  We use an IVR (Interactive Voice Response) to get audio content from the user group. Interactive Voice Response (IVR) is an automated telephony system that interacts with callers, gathers information and routes calls to the appropriate recipient. The user gives a missed call to a number and the the computer calls back the user asap and then the user can speak about the content and the system records it.

- **Moderation of Content**

  The process of moderation is decentralized in order to make it's working more efficient. It consists of two levels of decentralization - Community Representatives (Volunteers) and Content Moderators.

  1. Moderation by Community Representatives

     As soon as the audio file is recorded over IVR, all the community representatives (volunteers) receive a notification about the content in their phone and any of them can listen to it. The volunteers check if the audio file is empty or is it of very poor quality or if it's to abusive or politically motivated. Based on these, they pass only the relevant contents to the next level of decentralization i.e. content moderators.

  2. Moderation by Content Moderators

     Moderator looks at the contents passed on by community representatives. They basically perform following two tasks namely - selection of content and ranking of content. They do the selection based on various factors such as biasedness of the content, it's political impact etc. Accordingly they reject the contents

that are not suitable to be published. Then among the contents which are to be published, they rank them based on their understanding. The overall ranking of the content is determined with the help on a ranking algorithm which takes into account the ranks given to a content by each content moderator.

- **Publishing Content**

  The content which passed the moderation test are published in order of their ranking. The user can listen to content on demand, according to their preference. The user response i.e. whether a user listens to or skip a content is recorded and this information is used in the next step (Popularity Index).

- **Content Popularity Index**

  Based on the information of lifetime of the content viz-a-viz whether people are listening to the content or skipping it etc, a mathematical measure of popularity of content called Popularity Index is calculated. Based on the measure of popularity indices of contents of a particular topics, volunteers determine whether user likes to listen about the topic or not. This helps to get the information about the demands of user group and suggests volunteers to prefer contents on certain topics over the others.

- **Scoring System for Volunteers**

  It might happen that volunteers don't work efficiently and just randomly passes the content to content moderators. In order to check this, sometimes randomly some contents are sent both to volunteers and content moderators. If the decision of a volunteer about a content is same as the decision taken by the moderator, his/her score increases or else the score decreases. This is done with the help of a scoring algorithm. This score in turn decides the incentive that the volunteer gets for his/her work. So, in order to increase his/her incentive, the volunteer will try to increase his/her score and hence will work more efficiently.

5 Future work and Other directions

5.1 Future work

- We will look for devising the ranking algorithm which determines the overall ranking of the content taking into account the ranks given to a content by each content moderator.
- We will also to add certain features to our app by which the stories contributed by the community reporter are better organized based on the topics of the content.
- We will also look into some recent techniques for the scoring algorithm and Popularity Index which will be the major tools to maintain the loopholes in the system.

- We will also consider using more efficient vocabulary representation as compared to one hot encoding because vocabulary in this case would be very large.

5.2 Other Directions

- Audio Enhancement

    1. We also want to address the problem of poor audio quality. So we will try to bring in methods in the app to enhance the poor-quality audios.
    2. We will use digital sampling to reconstruct the audio as much as possible.
    3. We sample at least twice as fast as the highest frequency we want to record so that we can use Nyquist theorem perfectly reconstruct the original sound wave from the spaced-out samples.

- Speech Recognition and Automatic abuse filtering

    1. The user is asked to choose one of the provided languages so that it helps us to use a pre-trained neural network to convert audio to text.
    2. We will be using APIs like Watson speech to text or Google speech API for converting the audio to text.
    3. Depending on the abusive words in the language, we segregate the audios into abusive tolerable or abusive intolerable.

6 Budget and duration

6.1 Budget

No budget is required for this project.

6.2 Duration

We aim to complete this project by the end of the summer break i.e. the end of July, 2017.

References

[1] Aparna Moitra , Vishnupriya Das, Gram Vaani team, Archna Kumar , Aaditeshwar Seth*Design Lessons from Creating a Mobile Based Community Media Platform in Rural India, Published in Proceedings of the Conference on Information and Communication Technologies and Development on June 3 2016.*

[2] Mridu Atray, Aaditeshwar Seth *Reality Reporting and Moderation Apps for Community Reporters in Rural Areas, M.Tech Thesis submitted by Mridu Atray under supervision of Prof. Aaditeshwar Seth in July 2013.*

[3] Chris Peterson *A Brief Guide To User-Generated Censorship, Published on Jul.22, 2013, under general.*

[4] Cliff Lampe, Paul Resnick *Slash(dot) and Burn: Distributed Moderation in a Large Online Conversation Space, In Proceedings of ACM Computer Human Interaction Conference 2004, Vienna Austria.*

[5] Ceren Budak, Sharad Goel, Justin M. Rao *Fair and Balanced? Quantifying Media Bias through Crowdsourced Content Analysis, Published by Oxford University Press on behalf of the American Association for Public Opinion Research on 01 April 2016.*

[6] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon *What is Twitter, a Social Network or a News Media?, Published in Proceedings of the 19th international conference on World wide web Pages 591-600 Raleigh, North Carolina, USA — April 26 - 30, 2010.*

[7] Engin Bozdag *Bias in algorithmic filtering and personalization, Ethics and Information Technology Volume 15 Issue 3, September 2013 Pages 209-227.*

[8] Natali Ruchansky, Sungyong Seo, Yan Liu *CSI: A Hybrid Deep Model for Fake News Detection, In Proceedings of the 26th ACM International Conference on Information and Knowledge Management (CIKM) September 2017.*

[9] Roja Bandari, Sitaram Asur, Bernardo A. Huberman *The Pulse of News in Social Media: Forecasting Popularity, Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media.*