Automatic Moderation of User-Generated Content

Issa Annamoradnejad
Sharif University of Technology, Iran

Jafar Habibi
Sharif University of Technology, Iran

INTRODUCTION

In recent years, most popular websites, such as Facebook and Wikipedia, emphasize user-generated content, ease of use, participatory culture, and interoperability for end-users (considered as Web 2.0). These websites do not generate systematic content and only depend on user involvement and content generation for their popularity and growth. Due to the vast expanse of these systems in terms of users and posts, manual verification of new content by the administrators or official moderators is not feasible and these systems require scalable solutions. The current strategy is to use crowdsourcing, which usually consists of initial reports by the community on the activities of users and a final decision by the official moderators or experienced users. For example, in community question-answering websites, if a post is against the general rules of the system, other users can flag the post for its violation using a reporting system, which will be entered into a review queue for further processing.

The crowdsourcing strategy has serious problems considering the agility and high impact of new posts (Ipeirotis et al., 2010; Paolacci et al., 2010). The slow handling process of reports is the first problem that exists in all of these systems. In general, the community has to review or simply read a new post, notice its unlawful content, create a flag report for its unlawful content, and wait some time for moderators to review the report that is now in a reporting queue. This process is performed manually by moderators and users, a costly and timely effort that sometimes results in subjective and biased decisions. In addition, some content may never be reported by the community as they are shared privately inside a closed network or not noticed by a large number of readers. This could lead to the usage of platforms as a safe private place for illegal activity, such as terrorism. Finally, users may wrongly report the content of another user because of disagreements and add to the slow handling of reports or cause problems for the target user (because of an automated locking mechanism in case of excessive reports for a user).

Given the need to maintain user rules and the significant problems of crowdsourcing, providing solutions for automatic and fast detection of user violations can resolve the mentioned problems, save time and money, reduce decision subjectivity, increase content quality, and create a safe place for civil debate. In addition, the same automated models can be utilized to develop constructive recommender systems that would help users in new content creation or editing.

In this research, by addressing these problems of manual handling, the authors show the emerging need for automated moderation of user-generated content using the latest machine learning and data science methods. Some recent works that proposed case-by-case solutions will be reviewed and a novel taxonomy of moderation actions will be provided by collecting answers to a new questionnaire. In addition, the authors propose an automated system for recommending the type of required edits to improve

DOI: 10.4018/978-1-7998-9220-5.ch079

the content in a community Q&A website, such as Stack Overflow. Determining the type of required edits for a question would help the asker to fix the problems, reach more readers, and achieve answers to her questions. A more accurate question will help the readers to better understand the context of the problem and provide faster and more accurate answers to the question. Since the proposed approach only uses the question data and does not include previous user achievements or future community feedback on the question (such as upvotes and comments), it can be used as a recommender system for new users and question drafts. The model extracts features by three separate components of feature extraction, which will be fed to feature engineering steps. For the final classification task, the model is trained using a gradient boosting algorithm.

This chapter will present novel ideas to create a real-world recommendation system that can assist system users and moderators in identifying the existing issues of old questions, sharing new high-quality questions, reducing the time needed for performing moderation actions, and improving the overall quality of the system.

BACKGROUND

Previous studies proposed methods to automate a single moderation task in a specific context. Some recent examples include preventing the spread of false content across online social networks (Campan et al, 2017; Shrivastava et al., 2020), spam/ad detection in online encyclopedias (Green & Spezzano, 2017; Yuan et al., 2017), fixing tags (Stanley & Byrne, 2013; Singh et al., 2020; Khezrian et al., 2020), finding low-quality (Tavakoli et al., 2020; Selleras 2020) or duplicate questions (Wang et al., 2020) in community question-answering websites, and predicting punishment for toxic players of competitive games (Blackburn & Kwak, 2014). Majority of these cases use a supervised text classification model to separate unacceptable content or behavior from the rest.

Detecting cyber-bullying and hateful acts is another major concern in online systems, which requires quick and accurate detection and subsequent response. All users, even newly registered ones, can publicly share new posts and comments, which could result in a simple approach for cyber-bullying, racism, and other hateful acts. Several studies focused on the analysis, detection, and prevention strategies (Waseem & Hovy, 2016; Blackwell et al., 2017; Agrawal et al., 2018; Bugueno & Mendoza, 2019).

While the majority of user-generated content is text-based, most platforms allow sharing other types of data, such as voice-chats and gifs. In addition, some systems, like Instagram and YouTube, are completely based on non-textual content. Since user content rules generally apply to all types of data, detecting problematic content would require different machine learning models. Even though, a few studies addressed this challenge for certain types of content in recent years, e.g., moderation of voice-based communities (Jiang et al., 2019), the majority of previous studies on non-textual websites are using lexical and social features for the task (Chancellor et al., 2016; Liu et al., 2018). Non-textual content is more prone to copyright infringement, a study subject since the advent of these systems (Agrawal & Sureka, 2013; Brøvig-Hanssen & Jones, 2021).

In the last few years and with the recent advances in machine learning research, popular online social networks design and use private AI systems to automate a few user and content moderation tasks. These systems use machine learning classifiers trained on large corpora of texts manually annotated. Because of this method of learning, they suffer from bias in training (Binns et al., 2017), a new challenge that needs to be addressed. Recent studies also focus on developing explainable AI for moderation decisions based on AI-led systems to achieve transparency and accountability (Kou & Gui, 2020; Brunk et al.,