Mining the Web for Insights on Violence Against Women

JIHAD ZAHIR, HAJAR MOUSANNIF, ET AL., CADI AYYAD UNIVERSITY

Introduction

At least 37% of Arab women have experienced physical, psychological, economic, or sexual violence in their lifetime. Effective interventions to eliminate gender-based violence (GBV) rely on up-to-date data about personal attitudes and social norms. However, many of the traditional methods to measure attitudes and norms are costly, time consuming, and potentially biased. For example, self-reporting techniques are heavily affected by social desirability bias: respondents often provide answers that accord with social norms. Internet activity may be a useful source of information on social norms. Social media platforms in particular are important forums for people to express their thoughts, opinions, and sometimes even personal information and details of their daily activities. In this study, we develop Natural Language Processing (NLP) techniques to analyze and extract useful information on GBV-related attitudes from Arabic-language YouTube and Twitter posts.

Our Approach

Our analysis consisted of three major stages: 1) retrieval and classification of gender-related content; 2) sentiment analysis of the collected data and; 3) gender and geographic disaggregation.

1) Retrieval and Classification of Gender-related Content

We developed an algorithm to automatically classify gender-related YouTube videos. The algorithm was trained on a set of 600 videos, each of which was manually classified by: overall type (news/TV show, religious, adult, other); focus on conflict (stories about women in humanitarian settings); reports on acts of violence; whether created by civil society organizations, whether ironic in nature; and "polarity" (whether they exhibited positive, negative, or mixed opinions about gender equality). The trained algorithm was then applied to a set of 5,618 YouTube videos, and classified 90-95% of videos accurately.

Figure 1. The tool used by volunteers to annotate text.

2) Sentiment Analysis of the Collected Data
Sentiment analysis (SA) is the computational study of people's opinions, sentiments, emotions, and attitudes, as expressed in text. It provides a way to automatically evaluate and capture change in public opinion towards women and their rights. SA of Arabic poses some unique challenges given its structural and stylistic properties.

We tried two different approaches to analyze sentiments in YouTube comments and Twitter posts. First, we took two publicly available "dictionaries" containing lists of words annotated with polarity scores. We extracted the words in YouTube comments and Twitter posts, got their corresponding polarity scores from the dictionary, and then aggregated these scores to obtain the overall score of the initial text. We also considered the context in which a text appears — for example, whether the comment is associated to a video promoting GBV. Our second approach was machine learning-based sentiment analysis. We built a training dataset by asking volunteers to manually annotate thousands of text samples (Figure 1). Based on this training set, we built various classification tools and applied them to over one hundred thousand gender-related YouTube comments and tweets. Again, the classifiers performed very well, identifying nearly all features more than 90% of the time.

3) Gender and Geographic Disaggregation
Finally, we automated the disaggregation of text by sex and location. Textual sex-disaggregation relies on the assumption that, within a particular cultural context, stylistic differences exist between men and women. To build a training dataset, we used two name-sex inference databases and then built a model that predicted sex accurately in 90% of text samples. Our approach to automated location relied on the existence of geographically distinct dialects of Arabic. We were able to build models to accurately identify dialects for all regions except Gulf countries.

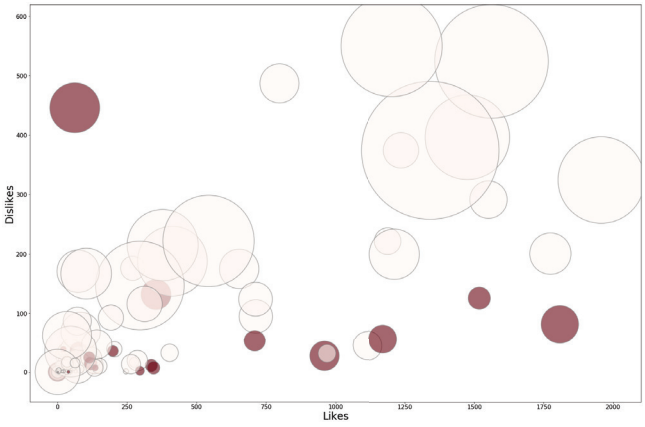
With respect to Twitter, we searched a gender-relevant keyword list to identify relevant tweets, and then applied our sentiment analysis algorithm to detect polarity. Twitter data allowed us to study relationships between positive and negative opinion holders.

These steps resulted in datasets of YouTube comments and tweets illustrating user attitudes about violence against women, classified according to sentiment polarity (positive or negative), sex of the user, and geographical location.

Results

Based on the articles defined in the Convention on the Elimination of All Forms of Discrimination against Women (CEDAW), we examined 12 topics in YouTube videos: women's work, women's education, beating women, female refugees, female genital mutilation, women in power positions, women in decision-making positions, women and politics, the eligibility of women to make decisions, women and leadership, and women and driving. Broadly, we found that negative videos were much more "liked" than positive ones. Many positive videos had high viewership (Figure 2), but the "likes to views" ratio was much higher for negative videos compared to the positive ones. "Likes to views" is a metric often used to measure popularity of videos in YouTube, and therefore we can conclude that negative videos are generally more popular. Negative sentiments were especially dominant in male-authored comments related to domestic violence and sexual harassment. We found that YouTube and Twitter sub-networks overlapped, representing ways by which both positive and negative opinions flow across populations.

Figure 2. Ratio of likes to dislikes for examined videos. The size of circles represents the number of views. Negative videos are colored dark, positive videos are light.



Implications

Automated mining of content on the World Wide Web is a cost-efficient way to learn about personal attitudes and social norms around VAW. To maximize the value of this information, data access, representativeness, and reliability should be more carefully considered.

Public APIs (application program interfaces) give access only to a small percentage of global data, and many provide access only to real-time data collection, making post-hoc analysis difficult. Moreover, APIs' usage conditions are continuously changing. These restrictions are understandable given the market importance of such data, but they represent serious barriers to research of the kind profiled in this brief.

The representativeness of online data is important for two reasons: does the online population represent the population as a whole and does the publicly available sample of web data represent the online population as a whole? As internet access becomes more universal, the first source of bias will ease and, in the meantime, can be quantified relatively straightforwardly. The second source of bias is not a major issue for YouTube, as search techniques can compile all videos relevant to a topic, however, it is problematic with respect to Twitter. Currently, researchers cannot assess the representativeness of the sample of free tweets made available through Twitter's API; more open algorithms would be useful.

Finally, the reliability of information on social media is a concern. Most false data is about demographic information (e.g., name, age, location); false opinions are less common. Sarcasm detection techniques are also rapidly evolving, allowing more nuanced analysis of user intent.

Computational approaches hold great potential for measuring attitudes and norms. In the short-term, this research group will create an interactive dashboard to make the usage of the tools described in this report more intuitive for all users. In the medium-term, we will add a forecasting tool to predict trends in attitudes and norms. We will also analyze how opinions flow across networks, as well as how to characterize the different kinds of personalities that hold positive and negative opinions about VAW.