Decision Support Systems xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

A system for intergroup prejudice detection: The case of microblogging under terrorist attacks

Haimonti Dutta[a,*], K. Hazel Kwon[b,*], H. Raghav Rao[c,*]

[a] Department of Management Science and Systems, University at Buffalo, 325 Jacobs Management Center, Buffalo, NY 14260, United States of America
[b] Walter Cronkite School of Journalism, Arizona State University, Phoenix, AZ, United States of America
[c] Department of Information Systems and Cyber Security, The University of Texas, San Antonio, TX, United States of America

ARTICLE INFO

Keywords:
Intergroup prejudice detection system
Machine learning
Logistic regression with regularization
Social media text classification

ABSTRACT

Intergroup prejudice is a distorted opinion held by one social group about another, without examination of facts. It is heightened during crises or threat. It finds expression in social media platforms when a group of people express anger, resentment and dissent towards another. This paper presents a system for automated detection of prejudiced messages from social media feeds. It uses a knowledge discovery framework that preprocesses data, generates theory-driven linguistic features along with other features engineered from textual content, annotates and models historical data to determine what drives detection of intergroup prejudice especially during a crisis. It is tested on tweets collected during the Boston Marathon bombing event. The system can be used to curb abuse and harassment by timely detection and reporting of intergroup prejudice.

1. Introduction

Prejudice is defined as "an antipathy based upon a faulty and inflexible generalization. It may be felt or expressed. It may be directed towards a group as a whole, or towards an individual because he is a member of that group" [1]. It is rooted in social categorization, by which a human simplifies the meaning of the social environment [2,3]. Social categorization forms an indispensable part of human thought and is therefore a precondition for expression of prejudice. Individuals perceive the social environment dichotomously, as "us" versus "others". Those who are not part of "us", are the so called out-group, and are perceived as less dynamic, complex, and individuated. Clashes of interests and values may occur among groups, but these *intergroup conflicts* need not be instances of prejudice. If realistic differences in interests and values (or intergroup conflict) causes *antipathy* it leads to intergroup prejudice.

Prejudice may exist in intergroup conflict, as Tropp concludes, "a single expression of prejudice ⋯ can have negative implications for intergroup relations" [4, p. 143]. That is, prejudice expressed on interpersonal level not only alienates the targeted out-group members but also encourages the development of dissent and negative behavior towards the whole out-group. Thus, *intergroup prejudice* is defined as a distorted opinion held by one social group about another, without examination of facts causing aversion, hatred and hostility. It is heightened during crises or threat and may lead to clashes of interests and values among groups. However we note that not all types of intergroup conflicts are instances of prejudice.

Tropp's remark is particularly pertinent in the context of social media, wherein prejudiced utterances are often expressed too carelessly, without thought on how other (dissimilar or divergent) group members would perceive them. This can lead to heightened sense of insecurity, anger and hostility. Unfortunately, a filtering mechanism – an editorial decision making process by which a particular message is selected, omitted, or revised before distribution to audience [5,6] – for prejudiced content is largely lacking in social media systems. This absence makes prejudiced messages spread much faster in online settings than in offline settings. A first step towards building such an editorial decision making process is to identify which messages express prejudice towards an out-group (also referred to as *intergroup prejudice*). This study builds a computational system for reliable detection of intergroup prejudiced cues in social media messages. While previous research has attempted to develop systems for rumors and interpersonal attacks [7-11], to the best of our knowledge, the problem of intergroup prejudice detection has not yet been explicated.

This paper is organized as follows: Section 2 reviews related work; Section 3 formally describes intergroup prejudice; Section 4 explains the utility of social media data, particularly Twitter; Section 5 describes the framework for detecting intergroup prejudice; Section 6 presents machine learning models and Section 7, the empirical results. Section 8 concludes the paper.

* Corresponding authors.
E-mail addresses: haimonti@buffalo.edu (H. Dutta), khkwon@asu.edu (K.H. Kwon), mgmtrao@gmail.com (H.R. Rao).

https://doi.org/10.1016/j.dss.2018.06.003
Received 30 September 2017; Received in revised form 2 June 2018; Accepted 13 June 2018
0167-9236/ Published by Elsevier B.V.

2. Related work

There is extensive research in social psychology examining the role of intergroup behavior and prejudice [2-4]. A socio-functional approach to intergroup prejudice [12] contends that humans are interdependent social animals thus evolved to maximize benefits of "group living" by effectively coordinating individual members into a "well-functioning group". In this process, individuals necessarily engage in vigilance to identify, minimize, and eliminate potential threats to collective living, such as threats to trust, group resources, and socialization systems. The detected threat is then displayed through high-arousal emotions such as fear, anger, and disgust. The socio-functional approach highlights that intergroup prejudice is an emotional product of the interplay between the characteristics of a target group and a given situation [12].

In a bid to study intergroup prejudice, we examined prior work that may fall in a similar domain as that of the current study: deception and fraud, use of offensive language and expressions of hatred. While these areas of work also pertain to the detection of anti-social messages, detection of intergroup prejudice is a problem differentiated from prior work.

2.1. Deception and fraud

The design of systems for detection of deception and fraud [13-17] has risen to prominence in recent years. Deception and fraudulent behavior can cause prejudice against the group to which that fraudulent individual belongs (for e.g. consumers may treat as out-group vendors who manipulate their reviews). However, since the goal of this paper is to identify prejudice in social media, it does not aim to look for cues pertaining to deception, slyness or treachery but focuses only on cues for prejudice.

2.2. Offensive language

The use of offensive language and hate speech by members of one group against another can provide cues for understanding dissent, hostility and resentment among groups. There have been a few systems designed for automatic detection of offensive language - the Smokey system[1] [18], can detect offensive comments; [19] describes an alternative method for flame detection; techniques that use more complex linguistic features for flame identification such as dependency structure analysis [20] and grammatical relations among words [21]; detection of offensive and non-offensive contents by exploitation of the lexical collocation of profanity [22] are some examples. Not every case of offensive language use is prejudice. However, when offensive language is contextualized in an intergroup relation, the probability of it being used in prejudiced expression may be high.

2.3. Hatred

Hate speech is one of the most obvious forms of prejudiced expression, and thus has the greatest resemblance to the current study. Warner and Hirschberg [23] present an approach to detect hate speech in online texts, where it is defined as abusive speech targeting specific group characteristics such as ethnic origin, religion, gender, or sexual orientation. In the context of social media, Kwok and Wang [24] build binary classifiers to detect anti-black tweets directed against blacks by employing labeled data from diverse Twitter accounts. More recently, Djuric et al. [25] propose an approach to the detection of hate speech in online user comments using a continuous Bag Of Words (BOW) neural language model. Apart from the existing hate speech detection studies, the current study develops the prejudice detection model by focusing on two aspects: (1) it captures additional cues beyond the use of offensive language (2) while hate speech detection tends to target a single group, for example, anti-semitism [23] and anti-Blacks [24], the current work examines comments against multiple groups in the context of a real world crisis event.

In sum, prejudiced messages in social media have a risk to go viral, and aggravate intergroup divides of the society. Detection of intergroup prejudice is a similar yet distinguishable problem from other anti-social message detection models. Specifically, the two premises that this study is grounded on are unique from other work. First, social media user interactions often engage multiple intergroup relations; Second, prejudiced messages include a broader range of expressions beyond offensive language uses. For the first premise, we develop a labeled data for multi-group cues. For the second premise, we add the emotional intensity measured via a sentiment analysis technique (such as [26] and [27]) as a relevant step to the detection of intergroup prejudice.

3. Intergroup prejudice under threat

Expressions of intergroup prejudice tend to become more intense than usual when society faces a collective threat such as unforeseen crisis (e.g. political crisis, natural disasters). During such an event, threatened individuals generate a large volume of information as an attempt to reduce uncertainties. A nontrivial portion of such information, however, is not credible, and even worse intends merely to attack or blame other social groups, and may often appeal convincingly to some audiences in spite of their suspicious veracity. A large part of intergroup prejudice literature discusses threat as a situational cause for prejudice to thrive. Accordingly, we propose several rules for detection of intergroup prejudice (denoted by R1⋯Rn) by referring to the literature on threat. We begin by pointing out the most basic "cognitive" element of intergroup prejudice – that the expression of prejudice must contain a target group cue [2]. A target group cue could be revealed in two ways, either as a group marker; or as an individual marker representative of the group.

R1. (a) Social group-indicative words and (b) individual names representative of a social group will appear more significantly in prejudiced messages than random.

If an indication of a target group is a cognitive dimension, behavioral and affective dimensions manifests the expression of prejudice [2] which becomes particularly salient when a community faces threat collectively. For example, macro-level social threats such as economic downturn, reduced social welfare, and terrorism, either elicited by specific entities or unspecified, are found to heighten interethnic prejudice [28]. Similarly, frustration-aggression hypothesis [29] suggests that threatened individuals release anxiety and reduce a feeling of powerlessness by putting others down [1,29,30]. That is, attributing responsibility for negative outcomes to nameable "scapegoats" help individuals restore personal control over their environment, and such an attribution process manifests through aggressive emotional expressions. According to socio-functional theory of intergroup prejudice humans maximize the benefits of "group living" for which group members are necessarily vigilant in identifying, minimizing, and eliminating potential threats to the collective living (such as, threats to trust, group resources, and socialization systems) [31]. Once a threat is detected, it is displayed through intensive emotional expressions such as fear, anger, and disgust. The intensive activation of emotion may sometimes accompany violent behavioral intention [2]. The literature suggests that aggressive behavioral markers and emotional accentuation should be more frequently found in prejudiced messages than in a random message. To develop the model features relevant to this behavioral and affective dimension, we propose the following linguistic cues as representations of verbal aggression and emotional accentuation:

[1] This system considers only insulting and abusive words in its "flame" detector but is equipped with a parser for syntactic analysis.