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,*
160 Jacobs Management Center,
University at Buffalo, Buffalo, NY, 14260.
[b] *Walter Cronkite School of Journalism,*
Arizona State University, Phoenix, AZ.
[c] *Department of Information Systems and Cyber Security*
The University of Texas, San Antonio, TX.

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.

Keywords: intergroup prejudice detection system, machine learning, logistic regression with regularization, social media text classification

[*] Corresponding author

Email addresses: haimonti@buffalo.edu (Haimonti Dutta), khkwon@asu.edu (K. Hazel Kwon), mgmtrao@gmail.com (H. Raghav Rao)

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 toward a group as a whole, or toward 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 amongst 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 $\cdots$ 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 amongst 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, 8, 9, 10, 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.

2. Related Work

There is extensive research in social psychology examining the role of intergroup behavior and prejudice [4, 3, 2]. 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.

Deception and Fraud: The design of systems for detection of deception and fraud [13, 14, 15, 16, 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.

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

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

prejudiced expression may be high.

Hatred: Hate speech is one of the most obvious form of prejudiced expression, and thus has the greatest resemblance to the current study. Warner et al. [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 et al. [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, 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, etc.). 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 $\cdots$ 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, 30, 29]) . 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:

R2: Verbal aggression represented by (a) the use of offensive words and (b) the use of violent behavioral cues is more likely to appear in prejudiced messages than in a random message.

R3: Emotional accentuation represented by use of all uppercase letters in certain words, is likely to be stronger in prejudiced messages than in a random messages. Conversely, intergroup prejudice is rooted in distancing oneself from a certain group; and thus opposite to sympathy toward other groups. Empathetic language use could be an inverse representation of prejudiced messages. Therefore:

R4: Empathetic expression is less likely to appear in prejudiced messages than in a random message. Even if we examine multiple group cues, it is possible that certain social groups occur more prominently than other groups in prejudice messages. Therefore, adding an "interaction" feature that gives weights to the verbal aggression used against prominent target groups could enhance the accuracy of detection.

R5: Mentions of prominent social groups are more likely to co-occur with verbal aggression markers in prejudiced messages than in a random message.

The prejudiced responsibility attribution does not ensure an examination of factuality. In fact, early public opinion scholars found intergroup prejudice frequently coupled with rumors during a crisis event ([1], [30], [32], [33], [34]). Public sharing of misdirected out-group blames via social media could produce undesirable consequences such as harming the reputation of certain groups or individuals, debilitating social trust, and aggravating group polarization or social exclusions [35]. By considering a prejudiced message as a type of unverified statement, we propose the rule that several fact-indicative linguistic cues, including the reference to news organization, addressable source of information (i.e., URL), numeric information, and breaking news, should be inversely associated with prejudiced messages.

R6: Fact-indicative linguistic cues such as (a) news organizational reference, (b) the use of URL, (c) numeric information, and (d) breaking news reference are less likely to appear in prejudiced messages than non-prejudiced messages.

Based on the rules above, and the incorporation of advanced machine learning techniques, we pose the following research question:

RQ1: What combination of rules and algorithms work best to detect intergroup prejudice from social media data with high precision and recall?

4. Data Description

Given the popularity, scope and speed in reaching the audience, as well as the real-world mobilization potential, social media are increasingly becoming the breeding ground for exploring social scientific queries such as public opinions, mobilization during political unrests, decision making and behavioral predictions. Twitter is one of the popular social reporting tools used extensively under a crisis or disaster situation ([11, 10]), and exhibits a mix of informative and undesirable messages. This study uses Twitter data collected immediately after one of the signature social crisis context in the U.S. - the Boston Marathon bombing incident. We chose a crisis context because the threat situation generates more prejudiced messages than in usual times. The search API was used with three keywords: #BostonMarathon, bostonmarathon, and boston. 30,951 users posted 68,087 tweets regarding the incident between April 15th – 29th, 2013. The number of retweets was 113,884.

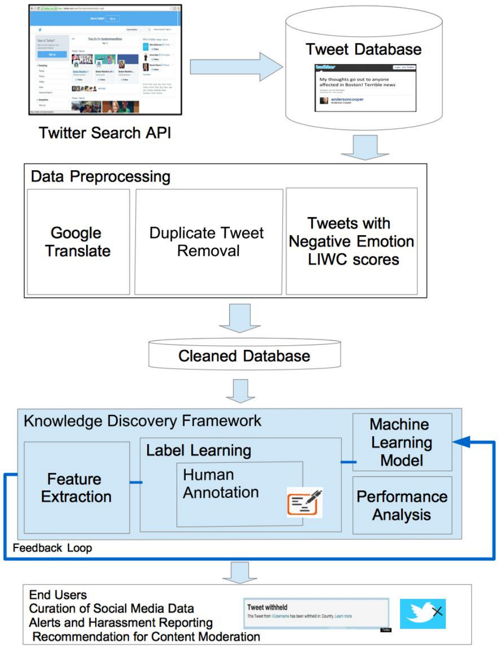
Twitter Search API
Tweet Database
My thoughts go out to anyone affected in Boston! Terrible news
andersoncooper
Data Preprocessing
Google Translate
Duplicate Tweet Removal
Tweets with Negative Emotion LIWC scores
Cleaned Database
Knowledge Discovery Framework
Feature Extraction
Label Learning
Human Annotation
Machine Learning Model
Performance Analysis
Feedback Loop
End Users
Curation of Social Media Data
Alerts and Harassment Reporting
Recommendation for Content Moderation
Tweet withheld
This Tweet from @Username has been withheld in: Country. Learn more

Figure 1: Knowledge Discovery and Machine Learning Framework for Intergroup Prejudice Detection

5. Knowledge Discovery Process for Intergroup Prejudice Detection

Figure 1 presents the components of the knowledge discovery framework for identification of prejudice from social media data. The key components include data preprocessor, feature engineering, and iterative learning: (a) *Data Preprocessor:* This component is involved with data cleaning including elimination of tweets with foreign language and removing words that occurred below the threshold frequency (= 1). More importantly, to enhance the accuracy of the model, we filtered out tweets with purely positive sentiments. Although text from social media may contain incorrect spellings, acronyms (for example, "gr8" or "gr8t" for the word "great"; "lol" for "laughing out loud") and short length (the length of Twitter message is limited to 140 characters), we do not perform spelling normalization. (b) *Feature Engineering:* Considerable effort was spent in designing proposed features which are described in more detail in Section 5.2. (c) *Label Generator:* No ground truth labels are available for the task described in this paper. To learn "subjective" concepts, human annotators provide an assessment of the labels for each Tweet based on words they contain. (d) *Iterative Learning:* The system described in this paper was carefully designed from the features using the labels to guide the process of learning. Manual tuning and detailed empirical analysis helped acquire the desired performance characteristics of the system.

5.1. Data Pre-processor

The following pre-processing steps are performed on individual tweets: First, we cleaned the data by translating[2] non-English tweets and removing duplicates. A big chunk of corpus contains replicas of the same messages. The number of tweets after removing duplicates was found to be 24,472.

[2] Google Translate was used for cleaning.

Second, we filtered out tweets with purely positive emotion - i.e., with zero negativity score - to enhance the accuracy of the model. The Linguistic Inquiry and Word Count (LIWC) software [36] was used to assign the sentiment scores by consulting a dictionary of words representing positive and negative emotion. It converts the usage of each of the single words within a tweet into either a 0 (not used) or a 1 (used one or more times) and then estimates the correlation between the occurrence of each word in the positive and negative emotion category with the sum of the other words in the same category By keeping tweets that showed at least some negativity, we narrowed the modeling scope down to detecting prejudiced messages out of negative messages corpus. This process was reasonable because prejudiced messages are within the domain of anti-social, negative messages. Excluding purely positive tweets indeed enhanced model performance.

Next, we eliminated stop words and special characters[3] from tweets using the dictionary (`https://github.com/mengjunxie/ae-lda/blob/master/misc/mallet-stopwords-en.txt`) in the Mallet software ([37]).

5.2. Feature Engineering

Features are extracted from the cleaned tweets stored in our corpus. Corresponding to the rules posited above, the following features are extracted for supervised learning (a) *Unigrams:* They are extracted after carefully removing occurrences of words with frequency 1 (since these are expected to contribute to noise in the data). All words are converted to lowercase unless the word was completely in uppercase in a tweet[4]. The total number of unigrams extracted for the study was 3567. (b) *Group indicative cues:* As the most basic cognitive unit

[3] If a word contains any of the following characters - ”.”, ”:”, ”!”, ”,”, ”?”, ”¨”, ”'”, ”)”, ”(”, ”¿”, ” ”, ”¡”, ”—”, ”=” at the beginning or end, those characters are removed.

[4] It is assumed that capitalization is an important property for identification of emphasis in text.

for prejudice to occur, the identification of a certain group must be revealed in a prejudiced message. To identify a group indicative cue, two human annotators[5] first verified whether a particular word was indicative of a group. Examples of such words include: immigrant, Russia, muslim, islam, Arab, Saudis, Republican, democrat. Annotators are trained a priori by a domain expert. The inter-rater agreement on the task was 97.52% (Cohen's kappa= 0.704).

In addition to the presence of group names, the annotators also reviewed the words to designate whether a word is the name of an individual who is representative of a certain social group group (e.g. representatives of a political party). This feature was an extension of the group feature, by considering an individual as a target for expression of prejudice. The inter-rater agreement on the task was reported to be 97.90% (Cohen's kappa = 0.755). (c) *Fact-indicative cues:* It is proposed that prejudiced messages contains unverified information such that fact-indicative features should be inversely related to prejudiced sentiment. The following features are identified as the fact-indicative cues: (i) *Presence or absence of URLs in tweets:* A feature was designed to test whether a tweet had an URL or not. On the one hand, the provision of URL may provide an external source to verify the statement. Considering that prejudice is a preconceived opinion[6] not based on actual experience, a tweet with a URL maybe unlikely to report intergroup prejudice. However, the provision of URL could be a persuasive strategy to make audience reinforce prejudice under the guise of seeming truthfulness. Thus, the presence or absence of an URL is an interesting feature for detection of intergroup prejudice. (ii) *Presence or absence of Twitter*

[5] The Stanford Named Entity Recognition (NER) toolkit `https://nlp.stanford.edu/software/CRF-NER.shtml` was first used to extract group and individual names from the tweets in our corpus. Only a small subset of names are identified; hence we resorted to human annotation. The Stanford NER could identify only 494 on the 1495 names we manually annotated, which is 33% of the total annotation we have.

[6] Oxford Dictionary

handles belonging to news and media organizations: A list of newspaper/press organizations are constructed based on LexisNexis search. This search retrieved 277 organizations, including domestic news media, major world news media, and international wire service. However, on closer examination it was found that some important newspaper names are missing (e.g., Boston Globe, Boston Heralds). To ensure the comprehensiveness of the list, additional information was obtained from the McClatchy-Tribune (M-T) partnership[7]. The final list included 932 organization names, after removal of repeated entries; also added are the broadcast network and cable news channel handles. Following this, Twitter handles associated with each organization name was found using the Twitter API[8] and a codebook containing 937 handles was created. The unigrams in each tweet are matched against this codebook to detect presence or absence of Twitter handles for news organizations.

(iii) *Presence of the word **BREAKING** in tweets:* A feature was designed to test whether a tweet had the word "BREAKING". It was proposed that the word "BREAKING" was primarily used in the context of urgent delivery of factual news therefore unlikely to be indicative of intergroup prejudice. (iv) *Numeric features:* A list of strings containing numeric characters was prepared as a fact-indicative feature component. The list includes phone number formats (e.g., nnn.nnn.nnnn; 01?nnnnnnn; 1-800-nnn-AAAA; 1-800-AAAA-AAA, +1-nnn-nnn-nnnn, with n referring to number and A referring to alphabet), time formats (e.g., n:nn; nn:nn), and currency format (e.g., *n.nn*).

(d) *Verbal aggression cues:* Two features are engineered to operationalize verbal aggression (i) *Presence of profane words:* A list of profane words was prepared by combining a few sources including online swearword collections

[7] M-T is the second largest newsgroup in the US. The largest group, Gannett News Service group, does not show the partnership list in the LexisNexis

[8] `https://dev.twitter.com/rest/reference/get/users/search`

online ("Banned Word List"[9]), the swearwords listed in LIWC dictionary, and the swearwords identified in the current text corpus. If the text of a tweet contains any one of these profane words, a binary feature has a value of 1 and 0 otherwise. In addition, we also incorporated the presence or absence of swear words (as defined above) as a feature - if a tweet contained any of the swear words, it has a value 1 and 0 otherwise. (ii) *Violent behavioral cues (Kill* features):* The most obvious form of prejudice often manifests itself by displaying a behavioral intention ([2]). Manual reading of the Boston bombing tweet messages also supports this idea in that hostile messages often contain the word "kill" and its variants (such as "killing"). Therefore, we created a binary feature (Kill* features) which is 1 if a tweet contained any occurrence of the word "kill" or its variants and 0 otherwise.

(e) *Emotional Accentuation:* Users often write out words or phrases with consecutive uppercase letters to emphasize their feeling, thoughts, and judgment. Note however, some words with uppercase letters are not necessarily associated with emotional or cognitive emphasis (e.g. #BOSTONMARATHON, RT, #TERRORISM, CNN). Accordingly, human annotators reviewed the uppercase words and created an exclusion list of 1,200 "conventional" uppercased words that are irrelevant to emotional accentuation. A binary feature was designed indicating whether a tweet contained one or more words in uppercase letters beyond the exclusion list. (f) *Features with empathetic connotations:* Considering that prejudiced sentiments are inherently intergroup divisive, empathetic emotions could be the opposite of prejudiced messages – for example, a majority of tweets could offer condolences to the victims and their families. Even if it could contain a negative emotion (sadness) and may designate a target group, the sympathetic message is far from prejudice. We consider that such

[9] www.bannedwordlist.com

words should have inverse relationship with prejudiced messages. For this feature, human annotators reviewed the bag of words to create the list of words[10] that connote either compassionate emotions or help-seeking intention. Based on the list, researchers reviewed the tweets that contained those words, and reduced to seven words that are clearly associated with sympathetic messages. Those words are help*, donat*, sadden*, heart*, thought, praying (other variations of pray* are often used in mixed sentiments and not included in this feature set), and tears.

Group/Indv. Name	#Pred.	#Non-Pred.	Total
ISLAM	30	71	101
MUSLIM	33	93	126
OBAMA	13	78	91

Table 1: Distribution of most frequently occurring Group/Individual names in our data.

(g) *Co-occurence Features between prominent social groups and aggression markers:* In natural language processing applications, n-gram based features (such as bigrams and trigrams) and parts of speech are often used to extract lexical patterns. Our initial datasets incorporated all bigrams and parts of speech based tags. However, it was found that this exponential increase in the number of features was not beneficial for improving precision and recall of the models. Consequently, we engineered "interaction features"[11] after carefully examining the text and keeping track of the frequency of occurrence of certain words associated with social categories. These are based on standard English usage <verb/adjective/noun> and occurrence of the name of a social category (group or individual name) in the tweet. In our dataset, the most frequently occurring (in at least 2% tweets) group/individual names are Muslim, Islam and Obama

[10] Available on request.

[11] This approach is motivated by the popular Brown clustering algorithm ([38, 39]) which is used extensively in natural language processing.

(Table 1 shows the frequencies in prejudiced versus non-prejudiced messages). Furthermore, the occurrence of the word "Muslim" (represented by the following unigrams - #muslim, MUSLIMS, muslims and muslim) often co-occurred with the following verbs(kill, killed, hate, hates, hated, hateful, #hate and shame) adjectives (stupid, evil, violent, VIOLENT) or nouns(terrorist, terrorists, #terrorists, suspect, suspects, suspected, suspects_, suspect's, #suspects, enemy, war, wars, #war and WAR). A set of binary features were introduced, depending on whether the tweet had the combination of group/individual name along with the corresponding[12] verb, noun or adjective. These were then subjected to the OR operation to generate one feature per group/individual name. This technique while requiring manual curation, helped reduce the bigram feature space substantially.

5.3. Label Learning

Each tweet in the corpus is assigned a label 1 if intergroup prejudiced sentiment is expressed. To generate labels for our learning framework, human annotators examined the tweets and labeled them manually. In the first round, two annotators split the data and identified independently whether a tweet contained prejudice. In the second round, a third annotator reviewed the identified tweets for reliability. The inter-annotator agreement between the first and second round reached 85.45% and 77.88% respectively. Only tweets identified as containing prejudice in both rounds are labeled positive - this accounted for 3.95% tweets in the corpus.

[12] The verb, noun or adjective can occur before or after the group or individual name

6. Machine Learning Model

The problem of prejudice detection from microblogs (such as tweets) is formulated as a classification task. Formally, given a training set S comprising of feature vectors $x_i \in \mathbb{R}^n, i = 1, 2, \cdots, N$ extracted from tweets and a variable y_i indicating whether the tweet has evidence of prejudice ($y_i = 1$), the goal is to find a function that discriminates the two classes (prejudice/no-prejudice)[13].

To solve the above problem, we model the class conditional densities $p(x|C_i)$ and the class priors $p(C_i)$ and then use them to compute posterior probabilities $p(C_i|x)$ using Bayes theorem. We use a logistic sigmoid function to find the posterior probability. This provides the baseline against which more sophisticated modeling approaches are evaluated. More specifically, the posterior probability for class C_1 can be written as: $p(C_1|x) = \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1)+p(x|C_2)p(C_2)} = \frac{1}{1+exp(-\alpha)} = \sigma(\alpha)$ where $\alpha = ln\frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)}$ and $\sigma(\alpha)$ is the logistic sigmoid function defined by $\sigma(\alpha) = \frac{1}{1+exp(-\alpha)}$. The posterior can also be written as a logistic sigmoid acting on a linear[14] function of the feature vector $p(C_1|x) = \sigma(w^T x)$[15] where w is the weight vector for the features and $p(C_2|x) = 1 - p(C_1|x)$. Occasionally, an intercept may be added, so that $p(C_1|x) = \sigma(\beta_0 + w^T x)$.

Let the vector $\mathbf{t} = (t_1, t_2, \cdots t_N)^T$ represent the vector of the predicted class probabilities for the training set S. Then, the cross-entropy error function for the prediction task is represented by $E(w) = -\ln\, p(t|w) = -\sum_{n=1}^{N} t_n \ln\, y_n + (1 - t_n)\ln\,(1 - y_n)$ Estimating the gradient of the function w.r.t. w yields $\nabla E(w) = \sum_{n=1}^{N}(y_n - t_n)x_n$. This expression takes the same form as the gradient of the sum of squares error function for the linear regression model and therefore a sequential algorithm can be used for learning the weights i.e. $w^{t+1} = w^t - \eta\nabla E(w)$.

[13] The class variable can have two values C_1 or C_2.

[14] We assume a linear model for the sake of simplicity.

[15] T represents the transpose of the matrix under consideration.

It was observed that simple logistic regression based models overfit the data easily and are unstable when the number of features exceeds the number of instances i.e. $n \gg N$. Consequently, the use of penalization techniques is investigated. The goal is to avoid arbitrary coefficient estimates by balancing the fit to the data and the stability of the estimates. The error function (shown above) can be penalized in the following way: $E^*(w) = E(w) - \frac{\lambda}{2}J(w)$. If an L_2 penalty term is used, such that $J(w) = \| w \|^2$ the technique is referred to as Ridge regression ([40]). While the accuracy of detection improved due to better bias-variance trade-off, it was observed that L_2 regularization did not necessarily yield sparse models. These models kept all the features and are incapable of generating small interpretable models.

A promising technique described in literature is the use of L_1 regularization (Lasso, [41]) which provides continuous shrinkage and variable selection. Although Lasso has been used extensively, it has several known limitations: (a) In the $n \gg N$ scenario, the Lasso selects at most N variables before it saturates, because of the nature of the convex optimization problem. It is also not well defined unless the bound on the L_1-norm of the coefficients is smaller than a certain value. (b) If there is a group of variables among which the pairwise correlations are very high, then the Lasso tends to select only one variable from the group and does not care which one is selected. (c) In cases where $N \gg n$, if there are high correlations between features, it has been empirically observed that the prediction performance of the Lasso is dominated by Ridge regression. The first two conditions above make Lasso an inappropriate choice in our setting. Our goal is to find a technique whose performance is at least as good as Lasso when $n \gg N$ and can also generate sparse models. Consequently, other techniques of regularization are tested including a linear combination of L_1 and L_2 regularization often termed as the Elastic Net ([42]). Formally, the regularization is achieved by: $E^*(w) = E(w) + \lambda[(1-\alpha)\frac{\|w\|^2}{2} + \alpha \parallel w \|]$ where α is a parameter that bridges the gap between Lasso (α=1) and Ridge (α=0). The tuning parameter λ controls the overall strength of the penalty.

This model provided the best test set accuracy while generating sparse interpretable models. To solve the above elastic net penalized logistic regression problem, a novel coordinate descent based optimization algorithm was adapted ([42]).

7. Empirical Analysis

The model performance is evaluated by using the following metrics – accuracy, precision, recall and area under the curve of a receiver operating characteristics graph.

7.1. Assessment of System Performance (RQ1)

7.1.1. Preparation of train-test data:

To evaluate the performance of the machine learning model(s), we built models on the train data and tested them on data not seen during the training phase (also called test set). The train and test splits are produced by stratified sampling of 3796 instances. 80% of the data representing prejudiced and not-prejudiced instances is used for training – this data set is called **train-Stratified**. The remaining 20% of data forms the test set.

7.1.2. System Building

The following method was undertaken to build the system: (1) For each instance in the training set, 3580 features are extracted. These include the 3567 unigram features and 13 additional features described in Section 5.2. Prejudiced messages are assigned a label 1. (2) A classifier is built on the training data. For baseline comparisons, logistic regression without regularization was built. It had the following parameters: (a) The weights of the classifier are learnt using a sequential algorithm described in Section 6. (b) The η parameter is set to 0.0019 after extensive grid search and experimentation with values in the range 0.0001 to 0.1. (c) The number of iterations of the sequential algorithm is set to 2000. (3) Next, penalized logistic regression models (L_1 (Lasso), L_2 (Ridge) and the Elastic Net) are built. The following details are relevant: (a) A ten fold cross-validation is adapted to select λ for Lasso and Ridge regression. (b) For the elastic net, the mixing parameter α is chosen by a grid search between 0 and 1, incrementing in intervals of 0.1. At each value of α, a ten fold cross-validation is run and the average error on the test set is recorded. The α at which the least training error is obtained is used for experiments. (c) To select an appropriate λ for the Elastic Net, the solutions for a decreasing sequence of values for λ are computed, starting at the smallest value λ_{max} for which the entire vector $\| w \| = 0$. The strategy is to select a minimum value $\lambda_{min} = \epsilon \lambda_{max}$, and construct a sequence of K values of λ decreasing from λ_{max} to λ_{min} on the log scale. The values chosen are $\epsilon = 0.001$ and $K = 100$. This scheme exploits warm starts and leads to a stable algorithm. For constructing the final model, we select λ_{min} provided by the above scheme. (4) Finally, the performance of logistic regression (with and without regularized) is assessed against three state-of-the-art text mining algorithms: (a) Support Vector Machines [43] - the **P**rimal **E**stimated **S**tochastic sub-**G**r**A**dient **SO**lver for SVM (Pegasos)[16] is used for experiments reported in this paper. (b) Random Forest [44] an ensemble of 50 decision trees grown on independently drawn bootstrap replicas of training data with the number of features selected at random for each decision split and (c) K-Nearest Neighbors with euclidean distance and number of neighbors set to 5. (4) The system is built on training data and tested on unseen data (20% of

[16] We use the linear kernel.

data with-held and not used for construction of system) and their performance reported using the metrics described in Section 7.

Method	AUC	Accuracy	Precision	Recall	FMeasure
Logistic Regression (No Regularization)	0.90	0.97	0.83	0.33	0.47
Logistic Regression + L_1 penalty (Lasso)	0.98	0.98	0.88	0.70	0.78
Logistic Regression + L_2 penalty (Ridge)	0.85	0.92	0.30	0.72	0.43
Logistic Regression + Elastic Net penalty	0.98	0.98	**0.91**	**0.70**	**0.80**
SVM (Linear Kernel)	0.73	0.46	1.0	0.46	0.64
Random Forest	0.63	0.27	1.0	0.27	0.42
K-NN	0.71	0.43	1.0	0.43	0.60

Table 2: Performance of logistic regression classifier(s) with and without regularization and three other text mining methods used in literature (Support Vector Machines, Random Forest and K-NN) on the test data. For the elastic net, $\lambda = 0.0009786$ and α=0.8; the Random Forest has 50 trees and 5 nearest neighbors are considered for K-NN.

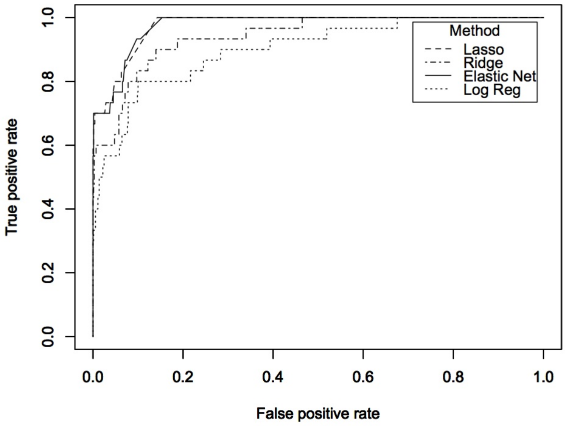

Figure 2: ROC curves for four methods used to model the data: Logistic Regression without regularization (baseline), Logistic Regression with Lasso, Ridge and Elastic Net penalties.

7.1.3. Results

The performance of the algorithm on test data is presented in Table 2. Logistic Regression penalized with elastic net penalty has both high precision and recall. The value of $\lambda = 0.0009786$ and α=0.8 for this model. It was observed that group indicative cues, along with offensive words are a clear indicator of prejudiced messages (Appendix A). Figure 2 plots the AUC curve for the four methods used to model the data. It was also noted that all three of the state-of-the-art text mining methods (SVM, Random Forests and K-NN) built high precision, but low recall systems due to overfitting of the training data.

Dataset	Train Set		Test Set	
	Pred	Non-Pred	Pred	Non-Pred
train-Stratified-60	90	2188	60	1458
train-Stratified-70	105	2552	45	1094
train-Stratified-90	135	3281	15	365

Table 3: Data set characteristics. “Pred” implies prejudiced; “Non-Pred” implies non-prejudiced

Dataset	Method	AUC	Accuracy	Precision	Recall	FMeasure
train-Stratified-60	Log Reg (No Regularization)	0.89	0.96	0.6	0.25	0.35
	Log Reg + L_1 penalty (Lasso	0.96	0.97	**0.66**	**0.55**	0.60
	Log Reg + L_2 penalty (Ridge)	0.96	0.54	1	0.02	0.03
	Log Reg + Elastic Net penalty	0.96	0.97	0.65	0.43	0.52
	SVM (Linear Kernel)	0.69	0.4	1.0	0.4	0.57
	Random Forest	0.62	0.23	1.0	0.23	0.38
	K-NN	0.69	0.38	1.0	0.38	0.55
train-Stratified-70	Log Reg (No Regularization)	0.88	0.96	0.71	0.22	0.34
	Log Reg + L_1 penalty (Lasso	0.96	0.98	**0.76**	**0.64**	0.70
	Log Reg + L_2 penalty (Ridge)	0.94	0.96	1	0.07	0.13
	Log Reg + Elastic Net penalty	0.95	0.97	0.75	0.62	0.68
	SVM (Linear Kernel)	0.69	0.33	1.0	0.33	0.5
	Random Forest	0.64	0.29	1.0	0.29	0.45
	K-NN	0.72	0.44	1.0	0.44	0.62
train-Stratified-90	Log Reg (No Regularization)	0.87	0.96	0.6	0.2	0.3
	Log Reg + L_1 penalty (Lasso	0.97	0.98	**0.92**	**0.73**	0.81
	Log Reg + L_2 penalty (Ridge)	0.95	0.96	0.32	0.67	0.43
	Log Reg + Elastic Net penalty	0.98	0.98	**0.92**	**0.73**	0.81
	SVM (Linear Kernel)	0.66	0.33	1.0	0.33	0.5
	Random Forest	0.67	0.33	1.0	0.33	0.5
	K-NN	0.69	0.4	1.0	0.4	0.57

Table 4: Performance of logistic regression classifier on the test data. The models are constructed by appropriately choosing $60, 75, 90\%$ of the intergroup-divisive tweets from the train set.

One thing to note is that proportion of prejudiced and non-prejudiced message can vary depending on a context. To simulate this scenario, we sampled different proportions of prejudiced tweets from the data - $60, 70$ and 90% and

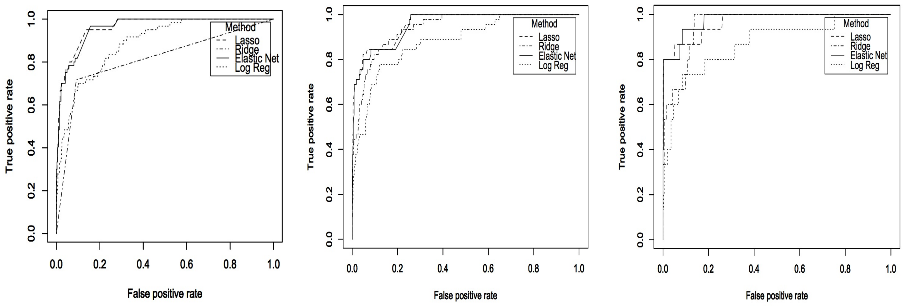

Figure 3: ROC curves depicting the performance of the models built by sampling $60, 70, 90\%$ data respectively.

studied whether the model(s) built are robust to these variations. Models train-Stratified-60, train-Stratified-70 and train-Stratified-90 are constructed (Table 3 presents the data characteristics) and the performance reported on the test set. Table 4 presents the performance of the models. The logistic regression models with L_1 penalty seem to have a reasonably good performance with both precision and recall in all the settings. The model with elastic net penalty follows closely and the differences in performances are not statistically significant. Elastic net performs as well as the Lasso model when the proportion of prejudiced tweets is very less in the testing environment. Figure 3 depicts the ROC curves for the models built using $60, 70$ and 90% of data. Addition of the penalty terms (L_1, L_2 and elastic net) significantly improved the performance over the baseline logistic regression models without regularization. The state-of-the-art text mining methods (SVM, Random Forest and K-NN) build low recall systems which is undesirable for intergroup prejudice detection. The logistic regression models are also robust to varying proportions of prejudiced and non-prejudiced messages which implies that they can be used in other online settings as well. This will be the focus of future work in addition to exploring other sources of data such as those obtained from Youtube, Reddit and similar websites.

Since the technique proposed for detection of prejudice (Logistic regression with Elastic Net) is a fairly new attempt, we presented performance of this method in three scenarios: (a) On related tasks – in particular, we examined the model performance on three public opinion data sets (b) Existing systems that detect antisocial messages (such as, hatred). (c) Comparison of human annotations versus state-of-the-art Named Entity Recognition (NER) systems. The results are presented in the Appendix. In all the cases, Logistic regression with Elastic Net had comparable or better performance than the state-of-the-art methods.

8. Discussion and Implications of the Work

Intergroup prejudice refers to a distorted judgement of another social group or representative(s) of the groups without examination of facts. Negative attitudes toward out-groups are then understood as an attempt to secure resources for the group-living. In crisis situation, behavioral and affective expressions of prejudice can worsen because such expressions could reduce uncertainty and anxiety in the minds of threatened individuals. That said, the prejudice expressed publicly via social media channels may pose a greater danger to today's pluralistic social living because social media is likely to spread such divisive messages, far more rapidly and broadly compared to private sharing. Large-scale prejudiced propagation online can bear qualitatively different implications from sharing on a private, small-scale interpersonal basis.

This paper presents a system for automatically detecting intergroup prejudice using machine learning. We develop a mechanism for collection of online data from social media, pre-process it by appropriately removing non-English text and duplicates; and extract meaningful features such as social group markers, verbal aggression markers, empathetic expression, and fact-indicative cues that are able to distinguish between intergroup prejudice and generally defined negative sentiment. The feature engineering was based on the rules inspired from social psychology theories of prejudice. Among them, some features (e.g., social group markers, verbal aggression markers, and the interaction feature) are introduced as prominent features for prejudice; while other features (e.g., empathetic expression, and fact-indicative cues) are included as inverse features. For the empirical case, human annotators labeled the prejudiced messages from the Boston Marathon bombing related tweets corpus. Only about 3% of the tweets exhibited prejudice when examined manually. The human-in-the-loop aspect of our system helps to provide better performance, however, it comes at a cost – such annotations may be difficult to obtain in practice, especially if the system is deployed for real-time intergroup prejudice detection. Our experiments revealed that automated NER systems can be used to generate features, although this may affect overall performance.

While the performance of the automatic system for detecting intergroup prejudice are satisfactory, some caution is warranted to prevent over-fitting. This is primarily due to the fact that the performance of the models are dependent on the quality of incoming social media data and its linguistic features. It is well known that limitations on the size of messages (such as 140 characters for tweets) causes users to type quick and short messages with many acronyms, spelling mistakes, emoticons and special characters that express special meaning. The use of spelling normalization and correction may help future research develop more robust models for detection of intergroup prejudice. Not withstanding these limitations, our models are able to identify messages that exhibit prejudice. To the extent possible, such messages can be alerted as having potential to spread misinformation and ill-will thereby assisting crisis information system managers, if needed. While it is possible to model the detection problem such that gradations of intergroup prejudice (for e.g. expression of contempt, sarcasm, hostile intent) are detected, this is left for future research.

It remains to be explored further, what implications the intergroup prejudice publicized via online networks may have on the social processes of crisis management. Negative consequences such as group polarization, diminished social trust, maladaptive reciprocation of online attacks, and even worse, spillover to hate crimes offline may be conceivable. At the same time, blaming certain social groups, for example politicians – can also be a violation of the First Ammendment. For exploration, it is necessary to understand the narrative characteristics and identify prejudice accordingly. While the manual detection of prejudiced information is a daunting task due to the sheer volume of social media messages, a machine-learning process can help assist detection.

9. Acknowledgements

The authors would like to thank Sujoy Dutta, Mimanshu Shisodia, Sneha Prabhakar, Sreya Nayak, Nitin Nataraj and Chris Bang for help with empirical analysis and annotation. This research is supported in part by the National Science Foundation under grants - SES-1227353, SBE-1419856, 1353119 and 1651475.

References

[1] G. W. Allport, L. Postman, An analysis of rumor, Public Opinion Quarterly 10 (4) (1946) 501–517.

[2] J. Duckitt, Prejudice and intergroup hostility, Oxford handbook of political psychology, 2003.

[3] G. W. Allport, The nature of prejudice: 25th Anniversary Edition, Addison-Wesley Pub. Co., 1979.

[4] L. R. Tropp, The psychological impact of prejudice: Implications for intergroup contact, Group Processes and Intergroup Relations 6 (2) (2003) 131–149.

[5] K. Barzilai-Nahon, Toward a theory of network gatekeeping: A framework for exploring information control, Journal of the American Society for Information Science and Technology 59 (9) (2008) 1493–1512.

[6] K. H. Kwon, O. Oh, M. Agrawal, H. R. Rao, Audience gatekeeping in the twitter service: An investigation of tweets about the 2009 gaza conflict., AIS Transactions on Human-Computer Interaction 4 (4) (2012) 212–229.

[7] R. Tian, Y. J. Liu, Isolation, insertion, and reconstruction: Three strategies to intervene in rumor spread based on supernetwork model, Decision Support Systems 67 (2014) 121 – 130.

[8] S. He, X. Zheng, D. Zeng, A model-free scheme for meme ranking in social media, Decision Support Systems 81 (2016) 1 – 11.

[9] K. H. Kwon, C. Bang, M. Egnoto, H. R. Rao, Social media rumors as improvised public opinion: semantic network analyses of twitter discourses during Korean saber rattling 2013, Asian Journal of Communication (2016) 1–22.

[10] O. Oh, M. Agrawal, H. R. Rao, Community intelligence and social media services: A rumor theoretic analysis of tweets during social crises, MIS Quarterly. 37 (2) (2013) 407–426.

[11] O. Oh, C. Eom, H. Rao, Role of social media in social change: An analysis of collective sense making during the 2011 Egypt revolution, Information Systems Research 26 (1) (2015) 210–223.

[12] C. A. Cottrell, S. L. Neuberg, Different emotional reactions to different groups: A sociofunctional threat-based approach to prejudice., Journal of Personality and Social Psychology 88 (5) (2005) 770–789.

[13] W. Zhou, G. Kapoor, Detecting evolutionary financial statement fraud, Decis. Support Syst. 50 (3) (2011) 570–575.

[14] N. Hu, L. Liu, V. Sambamurthy, Fraud detection in online consumer reviews, Decis. Support Syst. 50 (3) (2011) 614–626.

[15] N. Carneiro, G. Figueira, M. M. Costa, A data mining based system for credit-card fraud detection in e-tail, Decis. Support Syst. 95 (C) (2017) 91–101.

[16] M. S. Gerber, Predicting crime using twitter and kernel density estimation, Decision Support Systems 61 (Supplement C) (2014) 115 – 125.

[17] G. Ramesh, I. Krishnamurthi, K. S. S. Kumar, An efficacious method for detecting phishing webpages through target domain identification, Decision Support Systems 61 (Supplement C) (2014) 12 – 22.

[18] E. Spertus, Smokey: Automatic recognition of hostile messages, AAAI/IAAI (1997) 1058–1065.

[19] A. H. Razavi, D. Inkpen, S. Uritsky, S. Matwin, Offensive language detection using multi-level classification, Canadian Conference on Advances in Artificial Intelligence (2010) 16–27.

[20] A. Mahmud, K. Z. Ahmed, M. Khan, Detecting flames and insults in text, in: International Conference on Natural Language Processing, 2008.

[21] X. Zhi, Z. Sencun, Filtering offensive language in online communities using grammatical relations, in: Collaboration, Electronic messaging, AntiAbuse and Spam Conference, 2010.

[22] G. Xiang, B. Fan, L. Wang, J. Hong, C. Rose, Detecting offensive tweets via topical feature discovery over a large scale twitter corpus, ACM International Conference on Information and Knowledge Management (2012) 1980–1984.

[23] W. Warner, J. Hirschberg, Detecting hate speech on the world wide web, Second Workshop on Language in Social Media (2012) 19–26.

[24] I. Kwok, Y. Wang, Locate the hate: Detecting tweets against blacks., in: Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, 2013, pp. 1621–1622.

[25] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, N. Bhamidipati, Hate speech detection with comment embeddings, in: International World Wide Web Conference, 2015, pp. 29–30.

[26] B. Pang, L. Lee, A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts, in: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, ACL, 2004.

[27] N. Godbole, M. Srinivasaiah, S. Skiena, Large-scale sentiment analysis for news and blogs, in: Proceedings of the International Conference on Weblogs and Social Media, 2007.

[28] D. A. Butz, K. Yogeeswaran, A new threat in the air: Macroeconomic threat increases prejudice against asian americans, Journal of Experimental Social Psychology 47 (1) (2011) 22 – 27.

[29] C. I. Hovland, R. R. Sears, Minor studies of aggression: Vi. correlation of lynchings with economic indices, The Journal of Psychology 9 (2) (1940) 301–310.

[30] R. H. Knapp, A psychology of rumor, Public Opinion Quarterly 8 (1) (1944) 22–37.

[31] C. A. Cottrell, S. L. Neuberg, Different emotional reactions to different groups: A sociofunctional threat-based approach to prejudice, Journal of Personality and Social Psychology 88 (5) (2005) 770–789.

[32] R. L. Rosnow, Psychology of rumor reconsidered., Psychological Bulletin 87 (3) (1980) 578–591.

[33] S. Tamotsu, Improvised News: A Sociological Study of Rumor., The Bobbs-Merrill Company, 1966.

[34] R. H. Turner, L. M. Killian, Collective Behavior., Englewood Cliffs: Prentice-Hall., 1987.

[35] C. R. Sunstein, On Rumors: How falsehoods spread, why we belive them, what can be done, Farrar, Straus, and Giroux, NY., 2009.

[36] Y. Tausczik, J. W. Pennebaker, The psychological meaning of words: Liwc and computerized text analysis methods, Journal of Language and Social Psychology 29 (1) (2010) 24–54.

[37] A. K. McCallum, Mallet: A machine learning for language toolkit, http://mallet.cs.umass.edu (2002).

[38] P. Liang, Semi-supervised learning for natural language, Master's thesis, MIT (2005).

[39] J. Turian, L. Ratinov, Y. Bengio, Word representations: A simple and general method for semi-supervised learning, Annual Meeting of the Association for Computational Linguistics (2010) 384–394.

[40] A. E. Hoerl, R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics 12 (1) (1970) 55–67.

[41] R. Tibshirani, Regression shrinkage and selection via the lasso., J. R. Statist. Soc. B 58 (1996) 267–288.

[42] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Statist. Soc. B 67 (2) (2005) 301–320.

[43] S. Shalev-Shwartz, Y. Singer, N. Srebro, Pegasos: Primal estimated sub-gradient solver for svm, in: Proceedings of the 24th International Conference on Machine Learning, 2007, pp. 807–814.

[44] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[45] A. Conrad, J. Wiebe, R. Hwa, Recognizing arguing subjectivity and argument tags, in: Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics, ExProM, 2012, pp. 80–88.

Appendix A

Evidence is provided to verify that engineered features are able to differentiate between prejudiced and non-prejudiced messages. We perform Chi-squared tests to assess the quality of features and results are presented in Table 5.

Appendix B

In this section, we demonstrate the utility of the Logistic Regression model with Elastic Net regularization by analyzing its performance on the following three public opinion datasets with subjective labels.

(a) Argument Corpus: This corpus examines online editorials and blog posts concerning the debate over health insurance reform legislation in the United

Rules	Feature	Chi-square	p-value
R1	(a) Group	589.60	< 2.2e-16*
	(b) Individual	214.36	< 2.2e-16*
R2	(a) Profane Words	257.06	< 2.2e-16*
	(b) Kill*	1.21	0.27
R3	Emotional Accent	0.08	0.77
R4	Empathy	7.98	0.0047
R5	Co-occur: Muslim	46.90	< 7.45e-12*
	Co-occur: islam	139.60	< 2.2e-16*
	Co-occur: Obama	5.20	0.02
R6	News and Media Org	4.08	0.0432
	URL	0.08	0.77
	Number	0.42	0.52
	Breaking	0.07	0.79

Table 5: Chi-square tests for features. The asterisk next to the p-values marks the features that are capable of distinguishing between prejudice and non-prejudice.

States and annotates subjective arguments [45]. (b) Subjectivity Dataset: This data set contains subjective sentences (or phrases) collected from 5000 movie review snippets (e.g., "bold, imaginative, and impossible to resist") from `www.rottentomatoes.com`. (c) Polarity Dataset: This dataset is used to classify movie reviews as "thumbs up" or "thumbs down" using just the subjective portions of the document. It contains 1000 positive and 1000 negative reviews all written before 2002, with a cap of 20 reviews per author (312 authors total) per category. The Table 6 presents the results. Overall, Logistic Regression with Elastic Net Penalty has comparable (or better) performance than Logistic Regression without regularization on all three datasets discussed here.

Appendix C

In this section, we demonstrate the utility of the Logistic Regression model with Elastic Net regularization by analyzing its performance against a state-of-

Argument Corpus				
Method	Accuracy	Precision	Recall	FMeasure
Logistic Regression (No Regularization)	0.65	0.43	0.6	0.5
Logistic Regression + L_1 penalty (Lasso)	0.65	0.43	0.6	0.5
Logistic Regression + L_2 penalty (Ridge)	0.76	0.67	0.4	0.5
Logistic Regression + Elastic Net penalty	0.65	0.43	0.6	0.5
Subjectivity Data				
Logistic Regression (No Regularization)	0.61	0.63	0.56	0.59
Logistic Regression + L_1 penalty (Lasso)	0.85	0.86	0.85	0.85
Logistic Regression + L_2 penalty (Ridge)	0.83	0.84	0.82	0.82
Logistic Regression + Elastic Net penalty	0.85	0.86	0.84	0.85
Polarity Dataset				
Logistic Regression (No Regularization)	0.74	0.74	0.74	0.74
Logistic Regression + L_1 penalty (Lasso)	0.74	0.74	0.74	0.74
Logistic Regression + L_2 penalty (Ridge)	0.74	0.74	0.73	0.74
Logistic Regression + Elastic Net penalty	0.74	0.74	0.73	0.73

Table 6: Performance of logistic regression classifier(s) with and without regularization on three public opinion datasets.

the-art system[17] for detecting anti-social messages (such as hate and offensive speech). The authors collected tweets that contained terms from the Hatebase.org lexicon and labeled a sample of 25K tweets into three categories pertaining to hate, offensive speech or neither. A logistic regression with L_1 penalty was first used to select the best features and then L_2 regularization was used to construct the model. When this pre-trained hate speech / offensive language detector was run against our corpus, of the total of 150 prejudiced tweets in our corpus, the model from the state-of-the-art baseline was able to identify only 46 tweets. We have assumed in our experiments that hate and offensive language together comprise of the prejudiced messages while the neither label was selected as non-prejudiced label.

Appendix D

In this section, we demonstrate the effectiveness of using human annotations versus a state-of-the-art Named Entity Recognition (NER) system for generating

[17] https://github.com/t-davidson/hate-speech-and-offensive-language)

Split	Method	Type	AUC	Accuracy	Precision	Recall	FMeasure
90-10	Log. Reg. (No Regularization)	NER	0.76	0.97	0.91	0.53	66.67
		Org.	0.81	0.98	0.92	0.63	75
	Log. Reg. + L_1 penalty (Lasso)	NER	0.76	0.97	0.91	0.53	66.67
		Org.	0.81	0.98	0.92	0.63	75
	Log. Reg. + L_2 penalty (Ridge)	NER	0.63	0.96	1	0.26	41.67
		Org.	0.53	0.95	1	0.05	10
	Log. Reg. + Elastic Net penalty	NER	0.76	0.97	0.91	0.53	66.67
		Org.	0.79	0.98	0.92	0.58	70.97
80-20	Log. Reg. (No Regularization)	NER	0.75	0.97	0.68	0.5	57.69
		Org.	0.81	0.98	0.92	0.63	75
	Log. Reg. + L_1 penalty (Lasso)	NER	0.75	0.97	0.68	0.5	57.69
		Org.	0.78	0.98	0.94	0.57	70.83
	Log. Reg. + L_2 penalty (Ridge)	NER	0.62	0.97	1	0.23	37.84
		Org.	0.53	0.96	1	0.07	12.5
	Log. Reg. + Elastic Net penalty	NER	0.76	0.97	0.73	0.53	61.54
		Org.	0.78	0.98	0.94	0.57	70.83
70-30	Log. Reg. (No Regularization)	NER	0.73	0.97	0.63	0.48	54.32
		Org.	0.81	0.98	0.83	0.63	71.6
	Log. Reg. + L_1 penalty (Lasso)	NER	0.73	0.97	0.63	0.48	54.32
		Org.	0.81	0.98	0.83	0.63	71.6
	Log. Reg. + L_2 penalty (Ridge)	NER	0.61	0.97	1	0.22	35.71
		Org.	0.53	0.96	1	0.07	12.24
	Log. Reg. + Elastic Net penalty	NER	0.72	0.97	0.64	0.46	53.16
		Org.	0.79	0.98	0.79	0.59	67.5
60-40	Log. Reg. (No Regularization)	NER	0.69	0.96	0.6	0.39	47.06
		Org.	0.8	0.98	0.82	0.6	69.16
	Log. Reg. + L_1 penalty (Lasso)	NER	0.69	0.96	0.6	0.39	47.06
		Org.	0.8	0.98	0.82	0.6	69.16
	Log. Reg.+ L_2 penalty (Ridge)	NER	0.6	0.97	0.93	0.21	34.21
		Org.	0.55	0.96	1	0.1	17.65
	Log. Reg. + Elastic Net penalty	NER	0.69	0.96	0.59	0.39	46.6
		Org.	0.75	0.97	0.79	0.5	61.39

Table 7: Comparison of human annotation versus state-of-the-art Named Entity Recognition (NER) system using logistic regression classifier(s) with and without regularization on four different train-test splits of the dataset. Org. refers to the original model using human annotations and NER is the baseline Stanford NER system.

features. The Stanford NER system (`https://nlp.stanford.edu/software/CRF-NER.html`) was used to generate named entities and this was used in place of the Group/Individual features. The table 7 below presents the results. In almost all the cases, the human annotation decidedly outperforms the NER system except with L_2 regularization wherein low recall systems are built.