International Journal of Multimedia and Ubiquitous Engineering
Vol.10, No.4 (2015), pp.215-230
http://dx.doi.org/10.14257/ijmue.2015.10.4.21

A Lexicon-based Approach for Hate Speech Detection

Njagi Dennis Gitari[1], Zhang Zuping[1]*, Hanyurwimfura Damien[2] and Jun Long[1]

[1]*School of Information Science and Engineering, Central South University Changsha, 410083, China*

[2]*College of Information Science and Engineering, Hunan University, China*

gitaden2000@yahoo.com, zpzhang@csu.edu.cn, hadamfr@yahoo.fr, jlong@csu.edu.cn*

Abstract

We explore the idea of creating a classifier that can be used to detect presence of hate speech in web discourses such as web forums and blogs. In this work, hate speech problem is abstracted into three main thematic areas of race, nationality and religion. The goal of our research is to create a model classifier that uses sentiment analysis techniques and in particular subjectivity detection to not only detect that a given sentence is subjective but also to identify and rate the polarity of sentiment expressions. We begin by whittling down the document size by removing objective sentences. Then, using subjectivity and semantic features related to hate speech, we create a lexicon that is employed to build a classifier for hate speech detection. Experiments with a hate corpus show significant practical application for a real-world web discourse.

Keywords: *Lexicon, hate speech, subjectivity analysis*

1. Introduction

Most of the work on sentiment analysis focuses on review-based domains such as movie and product reviews. Currently, consumer generated content (CGC) through online forums, blogs and comment sections in news review sites are important sources of peoples' opinions on a variety of topical issues ranging from finance, education, religion, politics and a host of general social issues. The web discourse domain of sentiment analysis thus includes evaluation of web forums, newsgroups, and blogs. This domain has, however, become a common source of flames, virulent messages and rants. For instance, in the so called dark web forums, extremists and terrorist groups communicate, share ideologies, and use the forums as conduits for radicalization [1].

We are concerned with the task of determining hate speech sentiments in online forums, blogs and comments section in news reviews. A number of researchers have investigated the detection of flames and virulent messages in social media [2] as well as the spread of hateful messages in the dark web forums [2-3]. Different from other forms of rants or flames, hate speech uses offensive and threatening language that targets certain groups of people based on their religion, ethnicity, nationality, color or gender. The source of the hate messages is typically a member of a supposedly rival group such as belonging to another ethnic community. The dissemination of hate messages may be through dedicated web sites associated with a cohesive group of members but it may also be through popular sites such as Yahoo!, Twitter or Facebook where topical issues or news articles may elicit responses laced with stereotypical language. Lately, some works have introduced the domain of hate speech detection [4-5] but generally their focus is limited to supervised approaches to detection of racist speech. For our research, we abstract hate speech detection into three key target groups of race, nationality and religion.

ISSN: 1975-0080 IJMUE

Typical definitions of hate speech make reference to content of speech, tone of speech, an evaluation of the nature of that speech and targets of that speech as well as potential consequences or implications of speech act [6]. We reason that this sits well with sentiment analysis domain and envisages a design model that could capture the content and evaluative aspects of hate speech and to develop a system that can help in determining the severity of hate messages. For this task, we assume the documents to be topically relevant in respect to thematic areas mentioned above. Our work is concerned with the task of developing a classifier that can be applied to automatically detect the hate speech content in social media. Our main challenge is the lack of a labeled corpus that can be applied directly to this task. We use a rule-based approach for both subjectivity analysis and to develop hate speech classifier. Using sentence-level sentiment analysis techniques, we begin by subjectivity detection where we use rule-based approach to separate objective sentences from subjective sentences. Using subjectivity clues learned from the Multi-Perspective Question Answering (MPQA) [7] corpus and other sources, we train our rule-based subjective sentence detector. Since the clues have orientations to the opinionated, arguing, and polarizing topics, they are suitable for our sentimental based problem. To build our hate speech lexicon, we begin by extracting from the subjective sentences semantic word features that lender a sentence subjective. Since our domain of hate speech is heavily laden with domain dependence and context specific lexicon we need to augment our subjective semantic lexicon with corpus generated lexicon. Using bootstrapping and WordNet [8] we add into our lexicon, hate-related verbs and dependency-type generated grammatical patterns relating to the three thematic areas identified. Based on our lexicon, we create a hate speech detection application with three levels of: “No hate”, “Weakly hate” and “Strongly hate”. We test the application using our annotated corpus consisting of 500 labeled paragraphs. We summarize our process into the following four main steps:

Step 1: We use a rule-learning approach to extract subjective sentences.

Step 2: Using subjective sentences identified in step 1 above, we extract semantic and subjective word features.

Step 3: Using bootstrapping, we augment the lexicon in step 2 with noun patterns based on the semantic classes of religion, ethnicity and race and hate-related verbs.

Step 4: We build and test our classifier with our annotated corpus based on the features identified in Step 2 and Step 3.

The remainder of the paper is organized as follows. In the next section we introduce the domain of hate speech as it applies to web discourse. Then in Section 3 we review related literature on sentence level subjectivity detection and lexicon building. In Section 4 we introduce our hate speech corpus. In Section 5 we discuss our approach to subjectivity detection and lexicon building. We discuss our classifier development and experimental setup in Section 6 and 7. Finally, in section 8 we report conclusions and directions for future work.

2. Hate Speech

In the realms of social media, hate speech is a kind of writing that disparages and is likely to cause harm or danger to the victim. It is a bias-motivated, hostile, malicious speech aimed at a person or a group of people because of some of their actual or perceived innate characteristics [6]. It is a kind of speech that demonstrates a clear intention to be hurtful, to incite harm, or to promote hatred. The environment of social media and the interactive Web2.0 provides a particularly fertile ground for creation, sharing and exchange of hate messages against a perceived enemy group. These sentiments are expressed at news review sites, Internet forums, discussion groups as well as in micro-blogging sites.

According to Umati project [1], a project to monitor dangerous speech, and Hatebase[2], dangerous speech may be identified through the following ways:

(1) It is targeted at a group of people and not a single person. Dangerous speech is harmful speech that calls the audience to condone or take part in violent acts against a group of people. Thus, in the online space most common types of hate speech are related to nationality, ethnicity, religion, gender, social orientation, disability and class.

(2) May contain some of the hallmarks of dangerous speech, such as statements that compares a group of people with vermin or insects (metaphorical), suggest that the audience faces a serious threat or violence from another group and suggests that some people from another group are spoiling the purity or integrity of the authors' group.

(3) Dangerous speech often encourages the audience to condone or commit violent acts on the targeted group. The six calls to action common in dangerous speech are, calls to: discriminate, loot, riot, beat, forcefully evict, and kill.

For example, let us consider the following snippet attributed to the Israel's Prime Minister in 1988. It expresses, very strongly, negative sentiments against the Palestine people.

Example (1): "The Palestinians would be crushed like grasshoppers ... heads smashed against the boulders and walls."[3]

3. Related Works

3.1. Sentence-level Subjectivity Detection

A subjective sentence expresses some feelings, views, or beliefs. With sentence-level subjectivity, rather than individual words, each sentence in a given document is analyzed and checked to be subjective. When necessary, the subjective sentence can be further classified as being of positive or negative semantic orientation. Pang and Lee in [9] use a subjectivity detector to remove objective sentences from a given document. Then, using minimum cuts formulation, they integrate inter-sentence level contextual information with traditional bag-of-words features. They report considerable improvements over a baseline word vector classifier.

To learn subjective sentences Riloff *et al.* in [7] use two bootstrapping algorithms [10-11] to learn lists of subjective nouns from a large collection of unannotated texts. Then they train a subjectivity classifier on a small set of annotated data using the subjective nouns as features along with some other previously identified subjectivity features. A sentence is classified as subjective if it contains a subjective expression with a medium to high intensity otherwise it is classified as objective. This ensures only those sentences that are clearly subjective are classified as so. Besides identifying subjectivity and polarity of a sentence, [12] classify the strength of the opinions and emotions being expressed in individual clauses, considering clauses down to four levels deep. They leverage on a number of syntactic clues as well as subjectivity features tested in past research to recognize the subjectivity strength of a clause.

Ding *et al.* in [13] explored the idea of inter-sentential and intra-sentential sentiment consistency using natural language expressions. Instead of finding domain dependent opinion words, they showed that the same word could indicate different orientations in different contexts even in the same domain. Thus, they proposed to use aspect and opinion word pair to capture the context of the sentiment. Their

[1]http://www.research.ihub.co.ke
[2]http://www.hatebase.org
[3]http://whatreallyhappened.com/WRHARTICLES/palestinians.php

method thus determines opinion words and their orientations together with the aspects that they modify.

3.2. Lexicon Building

A number of previous works have attempted to generate sentiment words representing negative and positive orientation [14-16].The methods for generating opinion lexicon falls into two main categories, dictionary and corpus-based approaches. The former involves a static dictionary of semantically relevant words tagged with both a polarity label and semantic orientation score or reliability label [17-18]. The dictionary, in a number of the proposed methods, is initially generated using a bootstrapping strategy that uses a small set of seed opinion words and an online dictionary such as WordNet [8] and SentiWordNet [19]. There exist substantial resources of dictionaries of opinion lexicon built from mainly adjectives, but also from verbs, adverbs and nouns [17, 20]. Esuli *et al.* in [19] uses a semi-supervised method together with WordNet term relationships such as synonym, antonym and hyponymy to automatically generate a lexical resource that assigns each synset of WordNet three sentiment scores, summing up to one, regarding positivity, negativity, and objectivity, respectively. They leverage on a core seed of words that are known prior to carry a positive, negative or objective bias and iteratively add on new synsets using WordNet relations.

Dictionary based approaches generally suffer from the inability to find opinion words with domain and context specific orientations. Corpus-based approaches use a domain corpus to capture opinion words with a preferred syntactic or co-occurrence patterns. Using natural language processing rule-based techniques, syntactic, structural and sentence level features are used in determining the semantic orientation of words and phrases to be included in an opinion lexicon. With this method, a lexicon is populated with words and phrases that are more attuned to the domain by incorporating contextual features that could potentially change the semantic orientation of an opinion word. Features such as intensifiers could amplify or reduce the intensity of a neighboring lexicon item, while negations such as no, never, may change the directionality of a lexicon item. In [18] Wilson et al use a phrase-level sentiment analysis approach that first determines whether an expression is neutral or polar and then disambiguates the polarity of the polar expression.

3.3. Hate Speech Detection

A number of previous works are concerned with identifying racist hate groups, particularly against the black people. Social Network Analysis (SNA) techniques have been widely employed to identify web links between various hate groups. Chau *et al.* [21] uses analysis of hyperlinks among web pages to identify hate group communities. In [22], Zhou *et al.* use multidimensional scaling (MDS) algorithm to represent the proximity of hate websites and therefore to capture their level of similarity. Web mining and text mining techniques have been employed to analyze the contents of these web pages. To capture web contents they developed an attribute-based coding scheme with eight high-level attributes of communication, fundraising, sharing ideology, propaganda (inside), propaganda (outside), virtual community, command and control, recruitment and training. These are further refined into low-level attributes to capture low-level details such as telephone contacts, tactics, etc. Ting *et al.* in [23] use a combination of text mining techniques and social network analysis to locate and understand how various hate groups in Facebook share their ideas and attract new users. Using text mining techniques, they extract the keywords that are frequently used within the groups and rely on social network structure to find their relations. To extract the contents in keywords, they use TF-IDF

score, which they use to rank the feature vectors. Different from these studies, we are concerned with content analysis of hate speech using sentiment analysis techniques.

A semantic-aware approach for message filtering proposed in [24] uses grammatical relations among words to semantically remove offensive contents in a sentence. For each sentence, they begin by identifying offensive words, and then they extract semantic and syntactic relations among words in the sentence. Their approach is centered on what they term as modification and pattern integrity heuristic rules. However, their researches are concerned with offensive messages in general and not hate speech as in our case. One work that directly relates to hate speech uses labeled data to learn a Naïve Bayes classifier to distinguish between racist and non-racists tweets against black people [4]. The approach focuses exclusively on unigram features and is therefore a text classification application. Our approach is sentiment based that also considers the semantics of words by incorporating a lexicon of negative polarity words and noun patterns extracted using bootstrapping techniques. Perhaps, the closest to our work is in [5].Here, they present a supervised approach that categorizes hate speech by identifying stereotypes used in the text. Some of the categories identified include anti-Semitic, anti-Muslim and anti-African. They create a language model based on the anti-Semitic category and use correlation to identify the presence of hate speech in other categories. Different from their work, our research relies on a hate speech lexicon to develop classifiers and is unsupervised in orientation, it does not attempt at creating any categories on the basis of labeled data but only uses labeled corpus for evaluation purposes.

4. Hate Speech Corpuses

Our hate speech corpus consists of two different sources that have profoundly different orientation in terms of the target audience and presentation .For our main source, we crawled on diverse dates a total of 100 blog postings (documents) from 10 different websites, 10 for each site, from a list provided in the Hate Directory. This is a directory compiled by Raymond Franklin of sites that are considered to be generally offensive. One such site is Stormfront.org, a neo-Nazi's Web site, considered the first major domestic "hate site" in the United States. We were only concerned with blogs that were thematically related to areas of ethnicity, religion and nationality. Most of these sites provide theme-based discussion forums with numerous topical issues, essay contribution from members, invited guests as well comment section for the ordinary readers. Most of the discussions are intellectual discourses on subtle areas such as ideology and science. The text language used, though having some underlying bias, is rarely explicitly prejudicial, but rather contains a retinue of insinuations and innuendos against perceived rival groups. We refer to this as FIRST corpus. Our secondary source website consists of largely one paragraph snippets of quotes relating to the Israel-Palestinian conflict. The language used here is more direct and easily discernible even by a casual reader. Example 2 below is a quote, allegedly by Israeli Prime Minister Menachem Begin in 1982. We refer to this as SECOND corpus.

Example ***(2): "Our race is the Master Race. We are divine gods on this planet. We are as different from the inferior races as they are from insects. In fact, compared to our race, other races are beasts and animals, cattle at best. Other races are considered as human excrement. Our destiny is to rule over the inferior races. Our earthly kingdom will be ruled by our leader with a rod of iron. The masses will lick our feet and serve us as our slaves."***

We asked two graduate students in our university to undertake sample annotation representing approximately 30% of our corpus. For the FIRST corpus, we randomly selected 3 blog postings from each of the 10 websites for a total of 30 blogs. Similarly for

the SECOND corpus, out of a 150 page document, we divided the document into three sections of 50 pages each and selected the first 15 pages in each section for a total of 45 pages. To ease the annotation process, we asked that the ratings be done on a paragraph basis in case of essay-like documents and use snippets of sentences for the quotes. We rated using a 3 point scale of Not-Hateful (NH, both positive and neutral), weakly hateful (WH) and strongly hateful (SH). In total 180 paragraphs were labeled for the FIRST corpus and 320 paragraphs for the SECOND corpus. Table 1 below shows the agreement scores for the annotators, in both the FIRST corpus and SECOND corpus. The percentage agreement is 68%, but the Kappa (k) score was relatively low at 45%. To resolve cases of disagreement among our two annotators, a committee of the two and one author re-labeled the contentious paragraphs strictly on the basis of our definition of hate speech resulting in our gold corpus. In difficult situations a consensus was arrived at using a majority vote.

Table 1. Annotator Agreement for both the FIRST and SECOND Corpus

	NH	WH	SH	Total
NH	116	22	12	150
WH	36	56	30	122
SH	28	30	170	228
Total	180	108	212	500

5. Proposed Approach

Our task is to generate a lexicon of sentiment expressions using semantic and subjectivity features with an orientation to hate speech and then use these features to create a classifier for hate speech detection.

Our approach proceeds in three main steps. The first stage involves subjectivity detection, and is intended at isolating sentences that have subjective expressions from those that generally express objective sentiments. With hate speech, emotions, opinions, evaluations and speculations are laden with heavily subjective expressions. A direct benefit of using subjective sentences is to make detection of these opinions much easier. In the second stage we build a lexicon of hate related words using a rule-based method using subjective features identified from the sentences and semantic features learned directly from the corpus. Then in the final stage we learn a classifier that utilizes features created from our lexicon and use it to test for hate speech in a document.

5.1. Subjectivity Analysis

To learn a subjectivity classifier both rule-based and learning-based approaches have been used [25]. Rule-based methods do not involve learning and typically rely on a precompiled list or a dictionary of subjectivity clues. Learning-based approaches use Machine language algorithms on both labeled and unlabeled corpora to learn patterns or other subjective clues.

For the task of subjective sentence detection we employ a rule-based approach to classify sentences relying on a lexicon of well-established clues. In particular, we utilize two known sentiment lexicon resources of Wilson *et al*. [7, 26] and SentiWordNet [19]. The former comprises a list of over 8000 prior-polarity subjective clues tagged with positive, negative, neutral and both tags. Besides, the clues also have a reliability tag that labels each clue as either strongly subjective (strongsubj) or weakly subjective (weaksubj). In subsequent references, we refer to this lexicon as SUBJCLUE. We use the widely employed criterion that considers a sentence as subjective if it contains two or more clues designated as strong subjective clues. The algorithm is described in Figure 1. The latter lexicon, herein referenced as SWN, is built upon WordNet's synsets and

assigns to each synset a triple of negative, positive and objective score. We use SentiWordNet 3.0 that is aligned to WordNet 3.0. To determine whether a sentence is subjective or not, we extract the negative and positive scores for each sentiment word in the sentence. Then we subtract the negative score from the positive score to get the synset score. We sum the total score for all the words in a sentence and calculate their average. An average score of 0.5 and above lender a sentence as positive, while a score of -0.5 and below returns it as negative, otherwise it is determined to be objective. Therefore, sentences were determined as subjective if they are either negative or positive.

We begin by splitting the html documents into sentences, removing HTML tags in the process. Each sentence is consequently tagged using a Condition Random Field (CRF) - based POS tagger and the standard Penn Treebank tagset. This tagger assigns a part of speech to text tokens based on a sequence labeling method that assigns an appropriate POS label to every word token. Unlike the HMM the CRF avoids the label bias problem while at the same time enabling us to effectively disambiguate words that have multiple senses. We note that this is the only sense disambiguation we have attempted. Each token is reduced into a surface word and fed into the subjective classifier.

Not surprising that both corpora were relatively highly subjective. As Table 2 shows 56% of sentences in the FIRST corpus and 75% in the SECOND corpus were classified as subjective using SUBJCLUE lexicon. The SWN lexicon was the more conservative with 52% and 68% of the FIRST and SECOND corpus being subjective. However, interesting was that the SECOND corpus, though more subtle in presentation turned to be more subjective than the FIRST. We believe that though to a human reader the FIRST corpus seems more explicit, some of the words used do not naturally carry any sentimental value. We notice the use of words like a crocodile and grasshoppers, which, though having a neutral meaning in ordinary use are heavy with grammatical metaphor. We reason that to adequately capture the subjectivity of such words, their contextual use need to be put into consideration. In subsequent sections we use the subjective sentences resulting from the SUBJCLUE lexicon.

```
Predict subjSent  algorithm

Input: d: Text Document, sl: SUBJCLUE lexicon
Output: count: count of subjective words
//initialize list and count
Subjsentlist: list of subjective sentences
count←0
sentence ← “”
Begin
While (d! = null)
        sentence ← split (d)
        word← “”
        lex ← “”
                For each sentence ∈d
                        For each word ∈sentence
                        word←CRFtagger(word)
                          If word matches lex then
                              Count++
                                  If count >=2 then
                                    Output count
                                  End if
                                  addSubjsentlist(sentence)
                        End if
                         End for
                End for
End while
```

Figure 1. Subjective Sentence Prediction

Table . Subjectivity Results for FIRST and SECOND Corpus using SUBJCLUE and SWN Lexicons

	SUBJCLUES	SWN
FIRST	56%	52%
SECOND	75%	68%

5.2. Lexicon for Hate Speech

Based on the definition of hate speech in section II above, in the following we describe the process of extracting and developing a semantic dictionary of hate domain features from our corpus. The semantics of hate not only include typical opinion words with negative and positive polarities, but also employs rich linguistic stylistic devices. From similes to metaphors to juxtapositions, haters are in no shortage of language to pass their nihilistic motives. While a number of documents include the direct use of incitement and violent words, others use less explicit expressions and may be inexplicable if we rely only on opinion-oriented words. For example, rival groups are compared to concepts such as beasts, crocodiles and grasshoppers. The rule-based hate speech classifier that we ultimately create relies on three different sets of features. In the following we explain the features and the rules used to develop them.

(1) Negative Polarity

From the subjective sentences identified through the process in section V-1, we identify opinionated words that have a negative semantic orientation. All the word features in our corpus that match reliability tag of either weakly or strongly subjective and a polarity tag of negative in the SUBJCLUE lexicon are extracted and included in our polarity features lexicon.

(2) Hate verbs

As our second set of features we include “hate” verbs that are not part of the SUBJCLUE lexicon. The idea is to extract all the verbs that bear a relation with hate verbs from our hate corpora. Based on the definition of hate speech in section II, common in hate speech is the use of terms that condone and encourage violence acts. Such terms include the verbs discriminate, loot, riot, beat, kill, and evict. Beginning with an initial seed list of the six verbs, we use bootstrapping and WordNet’ssynsets, and hypernym relationships to build the list from all the verbs in the hate corpus and include as our hate lexicon only those verbs that overlap with our corpus but are not in the SUBJCLUE. If a verb in the seed list matches any of the verbs in the corpus list of verbs, then it is included in the lexicon list. For each round of iteration, we build a new seed list by reinitializing it with occurrences of unique words from the corpus list of verbs. The algorithm in Figure 2, captures the process of populating the hate verb lexicon using the seed list and the synsets and hypernym relationships.

Hate verb growing algorithm

Input: slist: An initial seed list of hate verbs, dv: A set of all verbs in the hate corpus
Output: hlex: A set of lexicon of hate verbs.
//initialize hlex with slist
hlex←slist
//create a set *s* and initialize
Set *s*←{}
For each word ∈slist
    s=Getsynset() and Gethypernyms()
        For each $s_i \in s$
            If(s_i appears in dv)
                Add s_i to hlex
            End if
        End for
End for

Figure 2. Hate Verb Lexicon Growing

The performance of this algorithm compares favorably with the classic Apriori algorithm for mining association rules. The main operations relate to acquiring a seed list for every iteration, getting from Wordnet the list of words that are semantically related to the seed list using the synsets and hypernym relationships and comparing them with the entire corpus of hate verbs. The time required for these operations depends on the number of words in hate lexicon and the size of seed list generated at each iteration. Thus the performance generalizes to O (n^2).

(3) Theme-based Grammatical Patterns

To generate grammatical patterns that depict hateful intent, a deeper representation of sentence or phrase structure that goes beyond typical surface lexical- syntactical structure was undertaken. Furthermore, to represent grammatical metaphors that depict the figurative meaning rather than literal meaning, constituent words cannot be applied directly to derive the meaning of phrases or sentences. Besides, using patterns to represent subjective expressions is more useful and richer than single words or arbitrary sized n-grams. Our task is to extract useful expressions related to our main themes of race, nationality and religion that can be used as lexicon in hate speech detection. We begin by extracting all noun words in our corpus relating to the three main themes of religion, race and nationality. We use the unannotated hate speech corpus described in section IV as the document from where relevant patterns are to be extracted. To create a list of seed words to help in generating patterns, we begin by manually adding from the General Inquirer (GI) [27] all the 15 words classified as race, referring to racial and ethnic characteristics as well as 72 out 103 words from the category ***relig***, which pertains to religious, metaphysical, supernatural or relevant philosophical matters. The 72 words were selected because they are nouns and uniquely appear in the lexicon. Further, with the support of Named Entity Recognition (NER) software, we identify sources and recipients of opinions that we include in our list of nouns. Some of the nouns that made to the list include names of nationalities such as Palestine and Israel as well as ethnic-religious groups such as Arabs and Jews. Then, using these words we extract grammatical relations that are structurally related to them using dependency type relations. We collect into our lexicon all governor-dependent pairs of words extracted from the corpus and incorporate them as co-occurrence two-word features. To reduce unnecessary features, a manual sorting is done to only retain features that are relevant to our task. To qualify as features for subsequent experiments our theme based patterns should have the following qualities:

a) The patterns should be constructed primarily from noun phrases

b) The extracted patterns should be semantic entities that can directly bear an opinion or are related to grammatical metaphors that can be used to express an opinion
c) Constituent words of the patterns should be directly linked to any of the three hate speech themes identified above

For grammatical analysis, an important step is to extract part of Speech (POS) information for each word in a sentence. Again, using Stanford tagger we present a Penn Treebank representation of all the words in a sentence reduced to their surface tag. For instance, using the tagger, the phrase "Money-Changers" in the following sentence is correctly recognized as a noun phrase.

"The Money-Changers have long since concluded that the main threat to their hegemony is Aryan man."

The/DT, Money-Changers/NNP, have/VBP, long/RB, since/IN, concluded/VBN, that/IN, the/DT, main/JJ, threat/NN, to/TO, their/PRP$, hegemony/NN, is/VBZ, Aryan/JJ, man/NN, /.]

Information extraction algorithms are commonly used to identify and extract relevant information from text using patterns of words and phrases. For example, AutoSlog [7], a weakly supervised pattern learning algorithm, uses a list of nouns as seed words and generates noun phrases whose head noun matches a noun in a document. Given a targeted constituent and a sentence, AutoSlog uses syntactic templates to heuristically identify an expression in the sentence that describes the conceptual role that the constituent plays.

Instead of using a phrase structure based parser like the AutoSlog for sentence presentation, we use a typed dependency structural representation that captures the individual words dependencies. In particular, we adopt the Stanford parser that models a sentence as a triple relation between a (governor, dependent) pair of words and a dependency type4 [28]. For instance, the sentence:

"Scurry around like drugged cockroaches" is resolved as follows using typed dependency.

[root(ROOT-0, Scurry-1), dep(Scurry-1, around-2), amod(cockroaches-5, drugged-4), prep_like(around-2, cockroaches-5)]

In this example amod (cockroaches-5, drugged-4); means that "drugged" serves as adjectival modifier of "cockroaches"; while prep_like(around-2, cockroaches-5); means that the like preposition introduces the noun cockroaches as a modifier to the adjective around. The dep (Scurry-1, around-2) suggests that a more precise dependency relationship is not directly discernible from the words available.

For our task, for every sentence, we generate dependency type relations that have a noun phrase (NP) as either a governor or a dependent. These relations includes nsubj type where NP is the syntactic subject of a clause, nsubjpass, a passive variant to nsubj where NP is the syntactic subject of a passive clause, amod ,described in the example above among others. For all the qualifying relations, we collect all the word pairs as two word co-occurrences and store them in a list. To ensure concurrence with the theme-based nouns, we map the noun list to the co-occurrence features and only those that match at least a member of the noun list are retained as features.

[4]http://nlp.stanford.edu/software/lexparser.shtml.

6. Aggregating Opinions for Hate Speech Detection

In the following we describe how we used the lexicon developed in the sections above to develop a rule-based classifier for identifying hate speech. Based on the lexicon created above, the features that are of interest to our application are typical negative opinion words described above as negative polarity and hateful words that are hate related but not part of the negative polarity lexicon. We present them as unigram one word features. Further, we have grammatical patterns inspired features that we capture as co-occurrence two-word features and represent context and domain-dependent features. To make a prediction as to whether a sentence is hateful or not, we use a sentence level assessment guided by the number and reliability of opinion words in a sentence as well as occurrences of other lexical expressions in the sentence. Key challenges are how to combine different features as well as to capture the context for the co-occurrence themed based features. Our incremental algorithm allows to vary the impact of the various sets of features by having hate lexicon and theme-based features depend on the semantic features. Each sentence that is classified as subjective from the annotated hate corpus is read and the base word in each sentence assigned a POS tag. Each POS tagged word is looked up in the various categories of lexicons. A sentence is classified into hate categories based on the following set of rules:

All negative opinionated words in a sentence are identified first. If two or more words tagged as strongly negative appear in a sentence we predict the sentence to be strongly hateful. Otherwise, if only a single word appears as strongly negative we predict the sentence as weakly hateful.

If only a single word in a sentence is tagged as strongly negative but one or more words appear in the hate lexicon then we predict the sentence to be strongly hateful. But if only one word appears from the hate lexicon without another from the semantic category, we predict it to be weakly hate.

If the governor-dependent pair of our co-occurrence patterns includes a theme-based noun and a word tagged as strongly negative or a word appearing in the hate lexicon then the sentence is strongly hateful. If the themed noun appears alongside a word designated as weakly negative we predict it as weakly hateful.

7. Experiment Setup

A system based on the proposed lexicon and algorithms described above has been implemented using Java programming language. This section evaluates the effectiveness of our system in predicting hateful sentences and rating their strengths. To accomplish this, our system consists of modules referencing different categories of feature sets as determined by the lexicon. A user can decide to use negative polarity features alone, negative polarity combined with hate features or alternatively a combination involving the two feature sets and additionally theme based features. We show that using the three categories assures the best prediction for hateful sentences.

The input to the system is a text corpus and a lexicon of subjectivity clues. After basic preprocessing, optionally, the textual corpus passes through a subjectivity analysis module and is transformed into a series of subjective sentences. Based on the annotation classification in section IV and the algorithms described in section V-2 we classify each sentence into either “Strongly hate”, “Weakly hate” or “No hate” categories.

Based on our incremental algorithm we carried experiments as follows:

In the first experiment we used the negative polarity clues extracted using the SUBJCLUE lexicon. For this experiment we use all the 3621 list of verbs, nouns, adverbs and adjectives in SUBJCLUE that are tagged with a reliability of strongly

subjective and a prior-polarity of negative. We used them to generate negative polarity lexicon as described in section V-2. A sentence was judged to be hateful if it contains at least two negatively tagged words, one of which must be strongly negative.

In the third experiment we included the "hate" verbs other than the ones in the SUBJCLUE lexicon list of strongly subjective. We rate their strength as equivalent to negative polarity lexicon in ii) above. In total we added 73 new verbs, including words such as kill and evict.

In our final experiment we included the lexicon of features derived from the theme-based grammatical patterns as described in section V-2 above. For this experiment we also include all the 1289 words in the SUBJCLUE that are tagged with a reliability of weakly subjective and a prior-polarity of negative. We extracted two word co-occurrence patterns satisfying the conditions in section V-2 on theme based features. Only patterns occurring with a frequency of at least two times in the corpus made it to the list. In total we extracted 103 such patterns.

7.1 Results and Evaluation

We used standard evaluation metrics for text categorization of precision, recall and F-score. Our evaluation scheme, like implementation, is incremental and focuses on the effects of the various sets of features. We base our evaluation on the total number of sentences classified under "strongly hate", "Weakly hate" and "Not hate" categories in our annotated gold corpus described in section IV. Precision is calculated as the ratio of the correctly categorized sentences to the total number of sentences classified under a particular category. Similarly, recall is calculated as the number of sentences that are correctly classified in a category divided by the total number of sentences that actually belong to that category. We use precision and recall for each category to calculate F-score using: 2(PR)/P+R. We compare our results with the annotated corpus described in section IV. The results in table 3 show the performance of our system in predicting "strong hateful" sentences. As shown in Table 3, using semantic features based on negative polarity alone the performance is below 70% in both recall and precision for both the FIRST and SECOND corpuses. However, when we include hate verbs, the results of precision increase slightly to above 70% for both the corpuses. We note the increase was more notable with the FIRST than the SECOND, this we believe can be explained by greater explicitness of FIRST corpus than the SECOND. The use of command and action oriented language is more associated with verbs than other forms of speech. However, the overall impact of the hate verbs lexicon was limited, largely because most the verbs generated are already part of the negative polarity lexicon. Further, we see some marginal improvements when theme-based features are included. .

To see the effectiveness of using subjective sentences, we compare the results by using all sentences (objective and subjective) from our annotated corpus. Like with subjective sentences, the processing extracts similar lexicon features as detailed in section V-2. The results show substantial drop in both precision and recall for both the FIRST and SECOND corpuses.

Past research on hate speech detection has focused mainly on determining racist texts in a document. In this work, we extend the domain of hate speech to include ethnicity and religion. Though comparison on ethnicity and religion aspects is infeasible due to lack of unifying dataset, we use the dataset used in [4] to compare the performance of our method with a supervised approach employed to predict racist tweets. Their approach employs a Naïve Bayes classification approach for the prediction of tweets as either racist or non-racist. We also include, for comparison purpose, earlier published results [5] derived from a supervised approach for

classification of anti-Semitic speech. As Figure 3 shows, our approach shows improvement in both precision and recall for the racist/non-racist class prediction.

Table 2. Classification Results for Strongly Hateful Sentences

Feature sets	FIRST Corpus			SECOND Corpus		
	Precision	Recall	F-score	Precision	Recall	F-score
Semantic	67.21	66.23	66.72	66.34	65.62	66.03
Semantic +hate	71.22	68.23	70.69	70.14	67.90	69.00
Semantic+hate+theme-based	73.42	68.42	70.83	71.55	68.24	69.85

Table 4. Classification Results for Strongly Hateful Sentences without Using Subjective Sentences

Feature sets	FIRST Corpus			SECOND Corpus		
	Precision	Recall	F-score	Precision	Recall	F-score
Semantic	58.42	61.12	59.73	56.68	57.54	57.11
Semantic +hate	63.24	64.42	63.82	61.56	62.24	61.90
Semantic+hate+theme-based	65.32	64.92	65.12	63.78	64.00	63.89

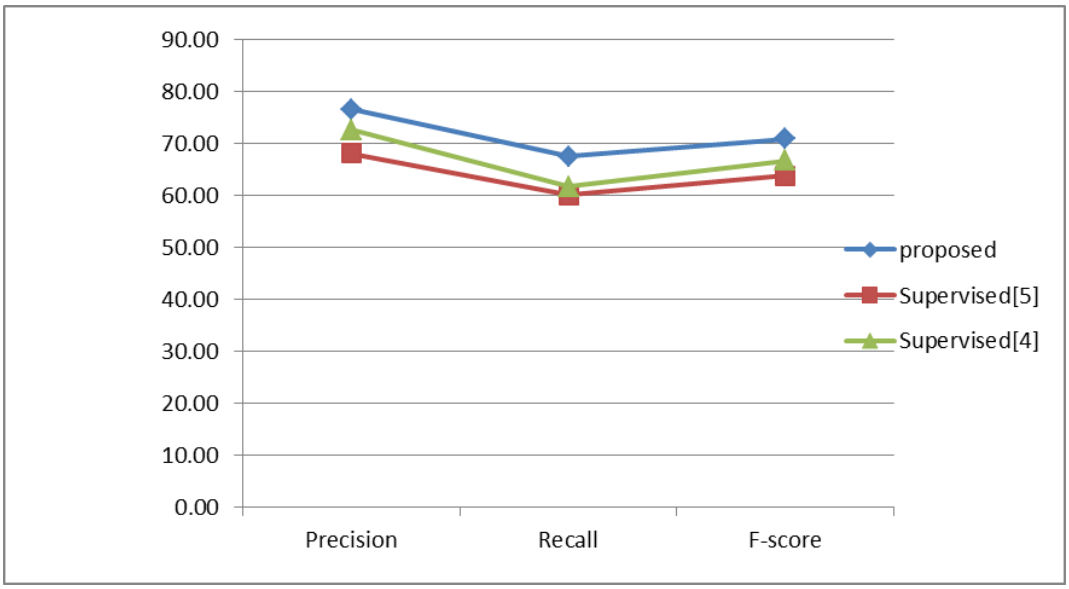

Figure 3. A Comparison of Precision, Recall and F-Score Results for the Racist/non-racist Classification; the Proposed Lexicon-based Method Produces Better Results in Both Precision and Recall

8. Conclusion

In this work we have investigated how dictionary and corpus based features can be combined to develop a classifier for hate speech detection. Beginning with a new corpus, a sentence-level test annotation was manually carried out on a representative

sample categorizing the hate corpus into three different categories. Then, using subjectivity analysis, objective sentences were separated from subjective sentences and removed from the corpora. A lexicon was created from semantic, hate and theme-based features and used in creating a rule-based classifier for hate speech detection. From the experiments conducted with lexicon extracted and evaluated on our sample annotation, the best results were achieved when we included semantic, hate and theme-based features. Further, the use of subjective sentences has been shown to improve on both precision and recall. In our future work we will incorporate topic-based approaches for identifying theme-based features as topics. Using topic models such as LDA, it could be possible to drop the assumption of topical relevance, hence making the application more widely applicable in web discourse domains. By expanding our annotated corpus, machine learning approaches such as SVM and maximum entropy can be applied directly with a possibility of increasing precision and recall scores.

Acknowledgements

This project was supported by the National Natural Science Foundation of China (Grant No. 61379109, M1321007) and Research Fund for the Doctoral Program of Higher Education of China (Grant No. 20120162110077). We would like to thank the anonymous referees for their helpful comments and suggestions. The authors are also grateful for the data and the help of Yuzhou Wang.

References

[1] A. Abbasi, H. Chen and A. Salem, "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums", (**2008**), ACM Transactions on Information Systems, pp. 1-34.
[2] E. Spertus, "Smokey: Automatic Recognition of Hostile Messages", Proceedings of the 8th Annual Conference on Innovation Application of AI (IAAI), (**1997**), pp. 1058–1065.
[3] A. Abbasi and H. Chen, "Affect intensity analysis of Dark Web forums", Proceedings of the 5th IEEE International Conference on Intelligence and Security Informatics**, (2007),** New Brunswick, NJ, pp. 282-288.
[4] I. Kwok and Y. Wang, "Locate the Hate: Detecting Tweets against Blacks", Proceedings of the 27th National Conference on Artificial Intelligence (AAAI), **(2013),** pp. 1621-162.
[5] W. Warner and H. Julia, "Detecting Hate Speech on the World Wide Web", **(2012),** Proceedings of the Workshop on Language in Social Media, Association for Computational Linguistics (ACL), pp. 19-26.
[6] C. Almagor and Raphael, "Fighting Hate and Bigotry on the Internet", Policy and Internet, vol. 3, no. 3, **(2011)**.
[7] E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions**", (2003) ,**Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP).
[8] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, "WordNet: an on-line lexical database", **(1990),** Oxford University Press.
[9] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts", **(2004)**, Association of Computational Linguistics (ACL).
[10] E. Riloff and R. Jones, "Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping", (**1999**), Proceedings of the 16th National Conference on Artificial Intelligence (AAAI).
[11] M. Thelen and E. Riloff , "A Bootstrapping Method for Learning Semantic Lexicons Using Extraction Pattern Contexts", **(2002),** Proceedings of the International conference on Empirical Methods in Natural Language Processing(EMLNP), Philadelphia, pp. 214-221.
[12] T. Wilson, J. Wiebe, and R. Hwa, "Just how mad are you? Finding strong and weak opinion clauses**", (2004)**, Proceedings of the National Conference on Artificial Intelligence (AAAI).
[13] X. Ding, B. Liu, and P. Yu, "A holistic lexicon-based approach to opinion mining**", (2008)**, Proceedings of the Conference on Web Search and Web Data Mining (WSDM).
[14] Y. Dang, Y. Zhang, and H. Chen, "A lexicon-enhanced method for sentiment classification: An experiment on online product reviews", **(2010)**, Intelligent Systems, IEEE, vol. 25, no. 4, pp. 46–53.
[15] Ounis, C. MacDonald, and I. Soboro, "Overview of the TREC-2008 Blog Track",**(2008),** Proceedings of the 17th Text REtrieval Conference, NIST.
[16] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis", **(2008)**, Foundation and Trends in Information Retrieval, vol. 2, nos. 1-2, Hanover, MA, USA, pp. 1-135.

[17] M. Taboada, A. Caroline and V. Kimberly, "Creating semantic orientation dictionaries", **(2006)**, Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), Genoa, pp. 427–432.
[18] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis", **(2005)**, Proceedings of the Human Language Technology Conference and the Conference on the Empirical Methods in Natural Language Processing (HLT/EMNLP), Vancouver, Canada, pp. 347–354.
[19] A. Esuli and F. Sebastiani, "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining", **(2006)**, Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC), Genoa
[20] V. Hatzivassiloglou and M. Kathleen, "Predicting the semantic orientation of adjectives", **(1997)**, Proceedings of the 35th Meeting of the Association for Computational Linguistics (ACL), Madrid, pp. 174–181.
[21] M. Chau and J. Xu, "Mining Communities and Their Relationships in Blogs: A Study of Online Hate Groups, (2007), International Journal of Human-Computer Studies, pp. 57–70.
[22] Y. Zhou, E. Reid, J. Qin, H .Chen and G. Lai, "US Domestic Extremist Groups on the Web: Link and Content Analysis", IEEE Intelligent Systems, vol. 20, no. 5, **(2005)**, pp. 44-51.
[23] I. Ting, S. L. Wang, H. M. Chi and J. S. Wu, "Content Matters: A Study of Hate Groups Detection Based on Social Networks Analysis and Web Mining",**(2013)**, Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Niagara, Ontario, Canada , pp. 1196-1201.
[24] Z. Xu and S. Zhu, "Filtering Offensive Language in Online Communities using Grammatical Relations", **(2010),** Proceedings of the 7th annual Conference on Collaboration, Electronic Messaging, Anti Abuse and Spam (CEAS), Redmond, Washington, US.
[25] S. Tan, Y. Wang and X. Cheng, "Combining learn-based and lexicon-based techniques for sentiment detection without using labeled examples", **(2008)**, Proceedings of ACM SIGIR Conference On Research and Development in Information Retrieval.
[26] J. Wiebe and E. Riloff, "Creating subjective and objective sentence classifiers from unannotated texts", (**2005**), Proceedings of the 6th International Conference On Intelligent Text Processing and Computational Linguistics , Mexico City.
[27] P. J. Stone, D.C Dunphy, M. S. Smith and D. M. Oglivie, "The General Enquirer: A computer".
[28] M.C. de Marneffe and C. D. Manning, "Stanford typed dependencies manual", (**2008).**

Authors

Njagi Dennis Gitari, He received his Bachelors of Education degree in Mathematics from Egerton University, Kenya in 2000 and a Master's degree in Computer Applications Technology from Central South University (CSU), China in 2004. He is currently a lecturer at Jomo Kenyatta University of Agriculture and Technology (JKUAT), department of Information Technology and a doctorate student at CSU. His current research interests include sentiment analysis and data fusion.

Zhang zuping, He received his Bachelors of Education degree in Mathematics from Hunan Normal University, China in 1985 and Master degree in Foundations of Mathematics from Jilin University, China in 1992 and PhD in Computer Applications Technology from Central South University (CSU), China in 2005. He is currently a professor at Central South University, School of Information Science and Engineering. His current research interests include information fusion and information systems.

Hanyurwimfura Damien, He received his Master's degree of Engineering in Computer Science and Technology from Hunan University in 2010. He is currently a PhD student at the College of Information Science and Engineering, Hunan University, China. He is also a Lecturer at the College of Science and Technology, University of Rwanda, Rwanda. His current research interests include information security and data mining.

Long Jun, He received his Bachelor, Master degree in Computer Science and Technology from Central South University (CSU), China in 1991 and 1994 and a PhD in Computer Applications Technology from Central South University (CSU), China in 2013. He is currently a professor at Central South University, School of Information Science and Engineering. His current research interests include network source management and dependable computing.

