Using Group Membership Markers for Group Identification in Web Logs

Jean Mark Gawron, Dipak Gupta, Kellen Stephens, Ming-Hsiang Tsou, Brian Spitzberg and **Li An**
gawron@mail.sdsu.edu
San Diego State University

Abstract

We describe a system for automatically ranking documents by *degree of militancy*, designed as a tool both for finding militant websites and prioritizing the data found. Our ranking system employs a small hand-selected vocabulary based on *group membership markers* used by insiders to identify members and member properties (*us*) and outsiders and threats (*them*). We use the same vocabulary to build a classifier. Evaluating several ranking systems by their correlations with human judgments, we show that the best ranker uses the small *us-them* vocabulary, outperforming one system with a much larger vocabulary, and another with a small vocabulary chosen by Mutual Information. We confirm and extend recent results in sentiment analysis (Paltoglou and Thelwall 2010), showing that a feature-weighting scheme taken from classical IR (TFIDF) produces the best ranking system; we also find, surprisingly, that adjusting these weights with SVM training, while producing a better classifier, produces a worse ranker. Increasing vocabulary size similarly improves classification (while worsening ranking). Finally, we experiment with adding *usage models* to both systems, models of how well each word's syntactic usage pattern matches the usage pattern in a class model; this model does not benefit ranking, but increases the precision of the classifier. Our work complements and extends previous work tracking radical groups on the web (Chen 2007; Zhou et al. 2007; Burris, Smith, and Strahm 2000), which classified such sites with heterogeneous indicators, including document, vocabulary, and morphological features. The method combines elements of linguistics, machine learning, and behavioral science, and in principle can be extended to data collection aimed at any group organized for collective action.

1 Introduction

For a variety of reasons, the problem of identifying web documents produced by particular groups has become a central concern for law-enforcement organizations, corporations, NGOs, social scientists, and public health agencies. This may be because a group espouses an agenda calling for large-scale social change (hate groups, terrorist organizations, political movements), or it may be that there is a particular issue around which members have joined after an appeal to collective action (the anti-vaccine movement, global warming, cancer and HMOs). In many cases the set of websites that promote a particular idea are only loosely linked together or not linked at all. Understanding what is going on with such groups – who they are attracting, what their successes are, where they are succeeding, their demographics – requires data. The focus of this paper is on how to collect such data, using the domain of white militant hate groups as an example.

The problem of finding relevant data is not trivial. Keyword searches on standard search engines turn up large numbers of false postives, because of the ambiguity of many crucial terms, because the sites are by their nature rare, and because they are not regarded as authoritative by the usual search engine ranking criteria (much hate group action business is conducted on blogrings). For example, a search on the keyword "ZOG" (Zionist Occupational Government) turned up Wikipedia pages (discussing hate group acronyms), game sites, commercial sites, and a zydeco band. A more fruitful approach, pursued on the pioneering Dark Web project as well as by others (Chen 2007; Zhou et al. 2007), is to crawl seed sites (provided by such resources as the ADL and SPLC websites and filtered search engine searches), download data, and do link and text analysis to find more data. This method can produce a lot of data; the Dark Web case study on U.S. domestic extremist groups reports 400,000 pages downloaded. The flip side of the problem of finding the data is how to sort through it when you've got it.

In this paper we propose to address the twin problems of finding the data and sorting it by pursuing two parallel approaches: In addition to building a straight militancy classifier, we also build a militancy ranker, a system that ranks pages by *degree of militancy*. One reason for this is that this is the right criterion of relevancy: The most militant documents are the most interesting. Another is that it provides a way of getting a handle on what may be massive amounts of data, one kind of problem encountered on the Dark Web project, and a general problem with web search (rarely is the issue too little data). A third reason is that it takes us one step down the road toward analysis. For example, GIS systems linking data to geographical location can use rank-

ing information to more accurately identify geographical hot spots (Tsou 2011). Our results indicate that the classifier and the ranker call for systems with significantly different properties.

The key challenge for both the classifier and the ranker is identifying randomly crawled pages as the products of known groups. This *group identification problem* is quite different from that of identifying a particular topic — a variety of subjects might be discussed on relevant pages, from music to electoral politics to online gaming. It bears a greater resemblance to the problem of identifying a particular author or a sentiment; what is needed is to find markers persistent across a variety of texts. When we add the issue of ranking by degree of militancy the problem becomes even more like one of the problems of sentiment analysis, since we are trying to locate a point on a scale.

The problem of tracking the sites of a particular group organized around a set of ideas is closely related to the problem of tracking the ideas themselves. Tracking specific ideas is difficult for the same reason text classification is: Texts expressing the same ideas may be expressed in a variety of ways. At the same time, texts expressing diametrically opposed ideas may be expressed in similar ways. This second difficulty is particularly salient when we bring into the picture opinion or ideology. Thus, we suggest focusing on group identification may provide information complementary to and beneficial to content information such as identifying a particular event or topic: knowing what's being said can be much easier if we know who's saying it.

On the one hand, focusing on the output of one strongly connected social group simplifies the natural language challenge by providing a more restrictive sublanguage. On the other, identifying the linguistic markers of a particular group also identifies a set of texts expressing a restricted set of ideas. This is particularly true of ideologically defined groups, but in the larger sense it is true of any cohesive group with a continuing social identity. Group members engaged in their group roles engage in a restricted set of conversations. This means that group identification and content identification can function as independent mutually reinforcing sources of information: Identifying a group can strengthen the likelihood of a particular content hypothesis; identifying the content can strengthen the likelihood of a particular group identification. Experience in the domain of sentiment analysis teaches us that where subtler linguistic discriminations are called for, multiple independent sources of information make for more reliable classification (Mullen and Collier 2004; Malouf and Mullen 2007; Prabowo and Thelwall 2009).

The experiments described below suggest that group identification itself is not that difficult as text classification problems go. We believe this is because groups organizing for collective action, bound together by an idea or group of ideas like the white militant groups in this study, have strong linguistic markers of their collective identity. Underlying this idea is the fact that individual relations (family, religion, ethnicity, friendships) are the strongest predictor of group affiliations and of collective action (Sageman 2004; Gupta 2008), and groups build up from such relations to ideas. In a very practical sense, group identification is prior to any conversion to a particular set of ideas; markers of group identity will thus be persistent features of group discourse, extending beyond manifestos actually engaged in articulating key group ideas, to forum interactions discussing group-particular music or games. More general social science motivations are outlined below (Section 2). The significant result is that ideas based on what we know about the psychology of groups can be integrated into a machine learning framework to produce an effective automatic ranking system.

The paper proceeds as follows:

1. Assuming a target group G whose texts we want to identify, we present a formal characterization of the problem as a **ranking problem**. Given a set of documents, we want to rank them on some (possibly continuous) scale
2. We present a **word-usage** model U incorporating syntactic features trained for a restricted vocabulary V. The group membership score of a document is a linear combination of the usage scores for the words in V. We present the usage models and several approaches to computing the weights of the linear combination.
3. We present a principle for extracting linguistic group markers by focusing on references to and properties of "us" and "them": the fundamental opposition that defines the group and the scope of its collective actions; we discuss a set of example markers derived from an analysis of a set of white militant data downloaded from active militant websites;
4. We describe 6 linear models: 3 Support vector machine models (SVMs) and 3 weighted feature models, and we train the models using 3 different feature sets, one of which is based on the white militant analysis in item 3; we show that this feature set yields the best ranker, but not the best classifier. Of the 6 models, 4 are usage based, 2 are not. We evaluate the models as classifiers and as ranking systems. The ranking evaluation consists of computing the correlation of the system rankings with those of a small set of human subjects given the same documents.

2 Group Identity Markers

The "rational actor" hypothesis, arguably the most widely accepted assumption in the social sciences, tries to explain patterns of human behavior as the natural result of individuals acting in their own individual interest. The idea is an organizing principle in disciplines as diverse as economics, artificial intelligence, psychology, and linguistics. Yet the growing field of social psychology starting, inter alia, with the seminal work of Tajfel (1978; 1981) is busily accumulating evidence of the importance of groups and group-identification in our decision-making process. These research efforts clearly demonstrate that our decisions are heavily influenced by the group(s) in which we claim membership. Our group or collective identity can even supersede our individual identity, in the sense that we may embark upon courses of action detrimental to our personal economic wellbeing, liberty, and life itself. This possibility is

strongest in groups in which the sense of self-identification is strongest. The key idea of this paper is that we can identify texts produced by members of such groups by identifying the markers of group identity in the texts.

Our starting assumption is that intense group identification requires a clear articulation not only of who "we" are, but also who"they" are — the outsiders, the other, the unbelonging, often, the enemies — an articulation that is central to all large-scale collective action from nationalism (Anderson 2003) to terrorism (Gupta 2008). In some cases a group is defined by a pre-existing language, but in most cases it is not; whether it is or not, an essential part of the process of dividing us from them is developing a group sublanguage. This may have a complex array of linguistic components, ranging from phonological to syntactic features, but an essential part of it is evaluative language referring to us and to them, as well as language referring to properties of us and properties of them. For well-established groups with a longer history the language includes a complex set of references to heroes, leaders, victims, and artists, as well as to subgroups, key events, key dates, and key writings and key works of art, including music and games.

Although group formation requires identification of "us" and "them," the mobilization of a large number of people for collective action requires a third factor: a clear articulation of an impending existential threat (Gupta 2008). This is not an unintuitive result. Behavioral research by the likes of Kahneman, Slovic and Tversky 1982 demonstrates the dominance of prospective loss over gains in our evaluation of uncertain futures: In brief, the prospect of losing what we have is a far more potent motivator than the prospect of gaining something new (Kahneman and Tversky 1979). As a result, from political extremism to electoral politics, fear tactics are a winning strategy. In accord with this idea, our us-them analysis will target language articulating threats as well as language referring to the enemy. In the white militant example, the in-group is members of the white race, the out group or enemies are the non-white population, including Jews and, depending on the group, the Catholics, but significantly, also a group of white people who are traitors to the race. The general existential threat is the degradation and pollution of pure white stock, but there are many more specific instantiations because degradation has many aspects.

We we will refer to the elements of the group sublanguage referring to us and them and to properties and products of us and them and to existential threats to us as the **us-them language**. Our hypothesis is that the elements of the us-them language are strong markers of group identity. Morever, the us-them language is largely learned, with more experienced speakers using it more fluently and more frequently. Speakers/writers who control a significant subset of this language are likely to be well-established in the group, and identifying a significant set of such markers in a text provides strong evidence of core group membership, in our example, a high degree of militancy.

As a first stab at implementing these ideas we took a representative sample of 74 militant web pages we judged to be extremely militant and extracted from them all proper names and all noun groups referring either to us-groups or them-groups including generic references to the groups as an entirety (*white people* or *ZOG*, for example). We also extracted Verb Phrases and Adjective Phrases referring to properties clearly identifiable as properties of us and them, as well as nominalizations referring to actions by us and them, and Noun phrases referring to threats or to symbolic or artistic products of us and them (such as *Collosians 2:8* and *African liberation flag*). Examples are given in Table 1.

Us	Them
Actions, products, threat	
be a hero	back gay marriage
be prepared	hate Jesus Christ
fight for white rights	deceive
protect your children	promote homosexuality
spread the good news	suppressing the truth
freedom	crimes against humanity
free enterprise	crooked banking system
hatred for the federals	cruise missiles
home schooling	cultural communists
core values	decay
death penalty	abortion
personal responsibility	African liberation flag
Collosians 2:8	AIDs plague
People, orgs	
Klansmen	ACLU
our revolutionary movement	CIA
leaderless resistance	ADL
our white brothers and sisters	ATF
Adolf Hitler	Bill Clinton
Apostle Paul	CBS
Ian Stuart Donaldson	Colin Powell
James Madison	Jesse Jackson
Branch Davidian Church	Jewry
family	democratic elite
folk	black community
forefathers	gays
heterosexual whites	liberals

Table 1: Sample White militant Us-Them phrases

3 The ranking problem

We formulate the problem of group identification as a ranking problem. Given some feature space $\mathbb{X}$, we seek a learning algorithm L that provides a **ranking function** γ that assigns a real number score to each element of $\mathbb{X}$:

$$\gamma : \mathbb{X} \rightarrow \mathbb{R}.$$

We assume a fixed set of classes $\mathbb{C}$ (which we will take to be $\{-1, 1\}$) and a set of labeled training documents $\mathbb{D}$,

$$\mathbb{D} \subset \mathbb{X} \times \mathbb{C}.$$

The ranking function γ is trained on D. That is, given D, the learning algorithm L produces γ.

Our assumptions about $\mathbb{X}$ are standard. A document d is represented as a vector $\overline{x}$ in $\mathbb{R}^m$. Then a set of test documents D of size n can be represented in a standard document matrix (Salton and McGill 1983; Manning, Raghavan, and Schütze 2008), an $m \times n$ matrix M, where column c_j is the m-ary vector for document d_j in D. We consider a class of linear models in which the n-ary ranking vector for the n documents in D is computed as follows:

$$s = w^{\mathrm{T}} \cdot \mathrm{D},$$

where w is an m-ary weight vector containing weights for each of the features. Thus learning a ranking function γ is learning the weights in w. For feature choice we again follow a standard assumption, the bag of words model: Each feature in a document vector represents a word. Thus m is the size of some fixed vocabulary V and feature choice is vocabulary choice. This admittedly highly restrictive formulation nevertheless allows us to explore a large class of linear learners as potential solutions to the ranking problem. In particular, for those classifiers which compute scores that can be interpreted as confidence levels in making a classification decision, the score can be taken to be our ranking score. Learning a ranking function thus reduces to learning a linear classifier. More particularly, since we use maximum margin classifiers (SVMs), the margin of a test example will be taken to determine "degree of militancy". As we will see below, this assumption is not unproblematic.

Under this formulation, designing a ranking system poses a cluster of related problems, including the choice of classifier, the choice of features, and the choice of document representation.

We focus here on the problem of vocabulary choice. Though vocabulary (or feature) choice is in principle less significant for a maximum margin model like an SVM, because learned weights can devalue less significant words, the problem of feature choice re-emerges once we train a ranking system rather than a classifier.

To train a ranking system, we could in principle train a system on data sorted into multiple classes. This is what is done in Koppel and Schler 2006 and Pang and Lee 2005. However, Koppel and Schler find considerable variation in performance of the resulting systems, depending on the dataset and the particular scheme of pairwise coupling, while Pang and Lee's approach requires a rating-scheme-specific similarity function. In addition, although human assignments of degree of militancy are strongly correlated (.81, $p < .0001$, see below), there are significant disagreements. It is much more difficult to get annotators to agree on what militancy score to assign documents than it is to get them to agree on whether they are or are not militant. We thus chose to train a standard binary classifier on data classified as either positive or negative (1 or -1) using its margin as our ranking score.

The downside of this approach is that such a classifier isn't necessarily being trained to learn degree of militancy. The more positively weighted features a document has, the higher its confidence margin, that is, the more evidence has been found for the classification; but high confidence in a militant rating is not the same as belief in a high *degree* of militancy. For that to follow on the bag of words model, it would have to be the case that a greater number of positively weighted militant words predicts a greater degree of militancy; that is an empirical hypothesis about the features being used, and as we shall see below, it is not true for every set of features: Using all vocabulary features or features selected by mutual information, a standard classifier which performs excellently on the classification task can perform quite poorly on the ranking task.

One approach, then, is to persist with the strategy of binary classification but to seek a set of features that does mark degree of group identification, that is, features such that their linear combinations ARE good indicators of degree of militancy. Our hypothesis is that if we focus on features that all have the property that they signal group identification (whatever they denote), more such features will reliably indicate greater militancy.

The final motivation for attending to feature choice is that it seems to be of help when we tackle tasks involving subtler discriminations like detecting sentiment or political orientation, where noise-eliminating strategies such as word and phrase selection based on semantic orientation or subjectivity have proven to be of help (Turney 2002; Mullen and Collier 2004; Whitelaw, Garg, and Argamon 2005). As we have seen, the us-them analyis chooses particular segments of the document that contain the strongest indicators of group identity.

4 Usage Model

Recently, the use of syntactic information has shown some promise in the field of sentiment analysis (Greene and Resnik 2009). The idea of using syntactic information in text classification tasks has a long history (Salton and McGill 1983) that has not always been crowned with success (Lewis 1992), for the same reasons, in part, that multiword features have been problematic in information retrieval: Single word features work better. It is easier to collect reliable statistics for single words; they cluster better by topic, and there are not as many of them. Our innovation here is to stick with single word features, but to factor the syntactic information into the weighting scheme.

The idea of the usage model is to measure how well the co-occurrence profile of a word in a test document matches its profile in a model by using word similarity measures as developed in work on distributional semantics (Schütze 1993; Grefenstette 1994; Dagan, Lee, and Pereira 1997; Lin 1998; Pantel 2003; Curran 2004; Zhitomirsky-Geffet and Dagan 2009; Turney and Pantel 2010). By the co-occurrence profile of a word we mean its co-occurrences with other words, in particular with the words that modify it and that it modifies. Our initial hypothesis is that a usage model will play a greater role in building a ranker than it does in building a classifier, because the usage model will be sensitive to the occurrence or non-occurrence of words in construction; thus a usage model can take into account co-occurrences like *white heterosexual*; where neither

word alone is a strong indicator of strong militancy, the co-occurrence is significant.

The usage model requires an alternative parallel representation of documents. Accordingly we assume a new space $\mathbb{Y}$, with document representations as assumed in distributional semantics: A document d_i is represented as an $m \times f$ **usage matrix** V, where m is the size of the vocabulary, as before, and f is the size of some set of syntactic context features. A set of n documents has n such matrices. We require some mapping from $\mathbb{Y}$ (where documents are matrices) to $\mathbb{X}$ (where documents are vectors), the space in which our ranking algorithm works.

In the document matrix for the document d_k, M^k,

$$\mathrm{M}^k[i,j]$$

contains a statistic about cooccurrences of the word w_i with the jth feature in document k. Thus $\mathrm{M}^k[i]$ is feature vector for the ith word in the document. We call such a vector a **word usage vector**. We compare the word usage vector of document d_k to its usage vector in a model, U. we assume U is of the same form as M^k, but $\mathrm{U}[i]$ is the feature vector for w_i computed from the entire positive training corpus. We map matrix M^k to its corresponding vector v^i in $\mathbb{X}$ as follows:

$$v^k[i] = \mathrm{sim}(\mathrm{M}^k[i], \mathrm{U}[i])$$

That is, $v^k[i]$ contains the similarity of vector $\mathrm{M}^k[i]$ to $\mathrm{U}[i]$. Since v^k characterizes the overall similarity of M^k to U, we call it a **conformity vector**.

We make the following assumptions about the statistics in the usage model U:

$$\mathrm{U}[i,j] = \mathrm{pmi}(w_i, f_j) = \log \frac{\mathrm{p}(w_i, f_j)}{\mathrm{p}(w_i)\mathrm{p}(f_j)}$$

That is, we take the relevant statistic about the co-occurrence of word w_i and feature f_j to be the **pointwise mutual information** of the word and feature in the training corpus as a whole. $M^k[i,j]$ uses the corresponding document probabilities:

$$\mathrm{M}^k[i,j] = \text{doc-pmi}(w_i, f_j, d_k) = \log \frac{\mathrm{p}^k(w_i, f_j)}{\mathrm{p}^k(w_i)\mathrm{p}^k(f_j)}.$$

We use cosine as our vector similarity measure. That is,

$$v^k[i] = \cos(\mathrm{M}^k[i], \mathrm{U}[i]).$$

As a word similarity model, this model closely resembles that of Pantel 2003. The novelty is that instead of using the similarity function to measure the similarity of two words, we use it to measure the match of the word's usage in a document to its usage vector in a class model.

To be concrete about the usage features in F in both U and M^k, we assume a **dependency model**. Features are pairs of grammatical functions and words. Thus, for example, in a large training corpus, tokens of the word *god* will occur with a number of different words functioning as *noun modifiers*. Here is a sample of such modifiers of *god* sorted by PMI value, normalized to give a unit vector.

```
NMOD:  dear:0.081   oh:0.080   almighty:0.077   creator:0.076
       almighty:0.068   omnipotent:0.067   lord:0.064
       immaterial:0.063   merciful:0.063   believed-in:0.060
       sky:0.060   extraterrestrial:0.060   thank:0.060
       praise:0.059   extra-worldly:0.059   vengeful:0.058
       sun:0.057   jealous:0.056   fertility:0.056   worship:0.055
       moon:0.054   israel:0.053   jeremiah:0.053   norse:0.053
       falcon-headed:0.053   thy:0.052   lion-headed:0.051
       within:0.051   doubting:0.050  elephant-headed:0.050
       undefined:0.050   !:0.050   whispered:0.049
       saves:0.049   good:0.049   ineffable:0.049 ...
```

The model and document vectors for *god* will require one position for each of these modifier statistics, as well as for many others. Building these vectors thus requires parsing both the training set and the test documents.

5 The experiment

We built 6 systems, testing them both as rankers and as classifiers, as well as testing them with 3 different feature sets. The 6 ranking models are the results of variation in two dimensions. In one dimension, we build models that have a usage model, versus a usage model combined with TFIDF weights, versus TFIDF weights alone. On the other, we built SVMs versus simple weighted systems that use linear combinations of similarity scores and/or TFIDF weights to compute a document score. To turn the simple scoring systems into classifiers, we chose a decision threshhold based on optimizing the F-score on the training set.

We describe the SVMs first. All the SVM systems were built using Joachim's SVM Light package (Joachims 2002).

1. SVM Sim: An SVM trained on document vectors v_k computed as described in Section 4. Thus for each word w_i in d_k,
$$\mathrm{v}^k[i] = \mathrm{sim}(\mathrm{U}[j], \mathrm{M}^k[j]) \quad (1)$$
$v_k[i]$ contains the cosine similarity of the distributional vector of w_i in d_k with the usage model distributional vector for w_i.
2. SVM Sim + TFIDF. Instead of $v^k[i]$ containing a similarity measure as in Section 4, we use:
$$\mathrm{v}^k[i] = \mathrm{TFIDF}(w_i, d_k) \cdot \mathrm{sim}(\mathrm{U}[j], \mathrm{M}^k[j]) \quad (2)$$
3. SVM TFIDF. Instead of $v^k[i]$ containing a similarity measure, we use:
$$\mathrm{v}^k[i] = \mathrm{TFIDF}(w_i, d_k) \quad (3)$$
For our TFIDF variant we use relative frequency times log of inverse document frequency:
$$\mathrm{TFIDF}(w_i, d_k) = \frac{\mathrm{count}(w_i, d_k)}{\sum_j \mathrm{count}(w_j, d_k)} * \log \frac{N}{\mathrm{doc\_freq}(w_i)},$$
where N is the number of documents. Thus, SVM TFIDF is a generic brand SVM document classifier.

The simple weighted systems all use the same document vectors as the 3 SVM systems, but rather than learning weights for the vector components and doing a weighted sum to compute a ranking score, we simply sum the vector components.

	$\mathrm{score}(v^k)$
Sim + TFIDF	$\sum_j \mathrm{TFIDF}(w_j, d_k)\mathrm{sim}(\mathrm{U}[j], \mathrm{M}^k[j])$
Sim	$\sum_j \mathrm{sim}(\mathrm{U}[j], \mathrm{M}^k[j])$
TFIDF	$\sum_j \mathrm{TFIDF}(w_j, d_k)$

These systems still fall within the linear model paradigm, though trivially. For example, in TFIDF, the generic IR system, the document independent weight for each word feature is just its IDF, and the document dependent feature value is just the relative frequency of the word, and in Sim, the weight for each feature is 1.

For all 6 system designs, we built ranking systems using 3 vocab sets as features:

1. The full vocabulary used in all our collected militant docs (full vocab), minus stopwords.
2. A vocabulary chosen by sorting the entire militant vocabulary by its mutual information with militant-class document and choosing the top 3500 words (MI vocab).
3. A vocabulary consisting of all the nonstop words that showed up in the phrases of our group-marker us-them analysis, minus stopwords (us them vocab).

For the usage model we parsed the training model with the Malt Dependency parser trained on Penn Treebank (Nivre 2003). A dependency DB was created, collecting dependency counts for each w, f pair and converting them to PMI values. Each f is in turn a pair of a dependency relation like SUBJ, OBJ, and NMOD and a word like *almighty*.

For the ranking experiment we collected 22 hand-selected sites ranging from totally non militant to militant with a sample of 3 web pages from each site. We had 5 human subjects rank them on a scale from 1 to 10 for militancy. This was done through a web interface, instructing them to rank websites promoting the superiority of the white race and violent means to achieve racial separation[1]. To evaluate our ranking systems, we had the systems compute militancy scores for each of the 66 pages, and then rank each website according to its highest scoring page (this seemed to be how our subjects judged militancy: one very militant page out of 3 was enough to rank a site highly). We then computed the average Spearman correlation rank coefficients of each of the ranking systems with the human subjects.

6 Results

System	Scores Vocabulary		
	Full	MI	Us Them
Sim	-0.13	-0.18	0.05
TFIDF	-0.03	-0.06	0.46
TFIDF + Sim	-0.44	-0.37	0.34
SVM Sim	-0.30	-0.28	0.01
SVM TFIDF	-0.10	-0.14	0.20
SVM TFIDF + Sim	-0.37	-0.41	0.05

Table 2: Average system correlations with human rankings. Average human-human correlation: .82

Table 2 shows the average Spearman rank correlation coefficients with humans for the 6 system designs and the 3 vocabularies. We used the Fisher transformation (Silver and Dunlap 1987) to average the pairwise correlations. For comparison, the average human to human correlation is also given. The 3 best systems, as far as matching human correlations, are the simple TFIDF system, the same system with a usage model, and the generic SVM classifier (SVM TFIDF) without a usage model, *all with the us-them vocabulary*. Clearly most of the work at capturing human militancy judgments is being done by the choice of vocabulary combined with the simplest TFIDF weighting scheme.[2] We speculate that the failure of the Sim model to measure up is due to the fact that the PMI computation alone loses information about relative feature frequency, and that this information is important to measuring degree of militancy; the Sim + TFIDF model remedies that problem.

Table 3 shows the results for the classifier comparisons for our 6 systems. The best classifier is the SVM TFIDF full vocab model, more or less the generic SVM. Moreover, the SVMs always outperformed the corresponding simple weighted systems as classifiers. This is not surprising, as SVMs have been shown to perform well in a number of document classification tasks. The benefits of using the largest feature set with an SVM are also unsurprising. Features that do little work or hurt in classifying have simply earned low weights. In contrast to the ranking task, the SVM Sim+TFIDF does show some promise: There is a considerable precision boost. This suggests the usage model may merit further investigation in applications calling for greater precision.

Turning to the non-SVM models, the Sim system is notable for its badness (Acc: 53.48); simply using the usage model scores yields close to random performance. The same model augmented with TFIDF scores (Sim + TFIDF) has much-improved accuracy (93.00), and outperforms TFIDF, the weighting system without Sim scores (89.13). This shows the usage model is contributing some information. However, the effect is much more muted among the SVM models, affecting only precision (for example, full vocab Acc: 95.43→90.88). We assume that what is going on is something like this: The raw TFIDF weights do not make a good classifier. The usage model bends them in the right direction; but the maximum margin computation bends them better, and is in fact hindered by noise that the usage weights introduce.

7 Discussion and Related Work

The most significant finding is that features hand-selected for their use in marking group membership, weighted only by their TFIDFs, made for the best ranking system, significantly outperforming another small feature set selected by Mutual Information. This provides strong evidence that the us-them analysis is turning up something significant. Admittedly the data set is small and we have yet to show that this set of features will extend robustly to ranking more diversified sets of documents, but we suspect that this feature set is a good seed. The importance of such *feature-engineering*

[1] `http://bulba.sdsu.edu/SWIDSAS/militant_eval_welcome.shtml`

[2] Only the differences between the positive and negative correlations are large enough to be significant.

Sim			
	Accuracy	Precision	Recall
Full vocab	53.48	33.68	90.28
MI vocab	70.94	48.70	77.78
us them vocab	79.63	71.88	63.89
TFIDF			
	Accuracy	Precision	Recall
Full vocab	89.13	90.28	90.28
MI vocab	76.81	65.00	72.22
us them vocab	68.49	47.62	83.33
Sim + TFIDF			
	Accuracy	Precision	Recall
Full vocab	93.00	95.65	91.67
MI vocab	91.05	92.42	84.72
us them vocab	86.15	87.88	80.56
SVM Sim			
Full vocab	89.42	93.10	75.00
MI vocab	91.64	95.16	81.94
us them vocab	89.52	90.77	81.94
SVM TFIDF			
Full vocab	95.43	92.00	95.83
MI vocab	94.19	90.91	97.22
us them vocab	91.64	87.84	90.28
SVM Sim + TFIDF			
Full vocab	90.88	100	83.33
MI vocab	91.46	98.39	84.72
us them vocab	91.17	98.39	84.72

Table 3: Systems used as classifiers

is well-known for a variety of applications, for example, in building good classifiers for spam detection and email filtering.

The other significant finding here was the clear separation between what makes a good classifier and what makes a good ranking system. The full vocab SVM TFIDF system was the best classifier; but the simplest possible weighted system (TFIDF only) with the hand-selected vocabulary made the best ranking system. Restricting the SVM to the hand-selected vocabulary made it much better than its full vocabulary cousin (-0.13 $\rightarrow$ 0.20), but still not as good as the weighted system with that vocabulary. The fact that TFIDF weighting played such a significnant role in the best ranker parallels the finding in Paltoglou and Thelwall (Paltoglou and Thelwall 2010) that various TFIDF weighting schemes gave better performance than binary features in sentiment analysis.

The poorer perfomance by the SVM in ranking suggests that the margins produced by the SVM weights do not correspond to degree of militancy, even though they do on balance seem to correspond to evidence for militancy. This finding is consistent with Koppel and Schler's 2006 finding, in the field of sentiment analysis, that binary-class training did not produce an SVM capable of correctly predicting neutral judgments.

A significant negative finding is that the usage model did not help with ranking. However, there is some evidence that it increased precision for the SVM classifier, which merits further investigation.

One extension based on related work seem natural: First, extending the feature set using mutual information may produce a vocabulary suitable for ranking more diversified data sets. The inspiration for this is Turney 2002, where choosing features with high mutual information for the words *good* and *excellent* produced results useful for sentiment analysis.

The SVMs evaluated here did not give the best rankers according to the current criteria, but they were not trained to. Despite this, the best SVM ranker placed second. More direct approaches to applying SVMs to ranking problems are pursued in Pang and Lee 2005 and Bickerstaffe and Zukerman 2010. A second future direction is to build classifiers of this form using an us-them featureset and a training set built from averaged human rankings and compare the results to the system described here.

Finally, this work follows the principle that combining information from multiple sources produces more reliable systems, a trend identifiable in a variety of prior work on tracking extremist groups. Abbasi et al. 2005 applies stylometry (authorship identification techniques) to the problem of identifying patterns of terrorist communication, which used both Arab and U.S. domestic extremist groups in their case studies. They used a heterogenous set of features, including document features like font and font color, as well as lexical and morphological features. Chau and Xu 2007 focus on the role of social network analysis in understanding the dynamics of links among hate group members, using among other data sets, data provided by Sageman 2004. Both bring content-external features of websites into the analysis. We conjecture that combining the approach pursued here with other independent sources of information can produce even better rankers.

8 Acknowledgments

This research was supported by NSF CDI grant # 1028177. The authors are also indebted to other members of the SWIDSAS research group, including Ick Hoi, Sarah Wandersee, Jennifer Smith, Ting-Hwan Lee, Amit Nagesh, Vickie Mellos. We are particularly indebted to Andrew McGladdery, who did the linguistic analysis that yielded the militant vocabulary, and to James Banker, who helped with the design of the pilot study of the human evaluations.

References

Abbasi, A., and Chen, H. 2005. Applying authorship analysis to extremist-group web forum messages. *Intelligent Systems, IEEE* 20(5):67–75.

Anderson, B. 2003. *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. New York: Verso.

Bickerstaffe, A., and Zukerman, I. 2010. A hierarchical classifier applied to multi-way sentiment detection. In *Proceedings of the 23rd International Conference on Computational Linguistics*, 62–70.

Burris, V.; Smith, E.; and Strahm, A. 2000. White supremacist networks on the internet. *Sociological Focus* 33(2):215–234.

Chau, M., and Xu, J. 2007. Mining communities and their relationships in blogs: A study of online hate groups. *International Journal of Human-Computer Studies* 65(1):57–70.

Chen, H. 2007. Exploring extremism and terrorism on the web: the dark web project. In Yang, C. C.; Zeng, D.; Chaur, M.; Chang, K.; Yang, Q.; Cheng, Z.; Wang, Jue annd Wang, F.-Y.; and Chen, H., eds., *Intelligence and Security Informatics: Pacific Asia Workshop, PAISI 2007*. Berlin: Springer. 1–20.

Curran, J. 2004. *From Distributional to Semantic Similarity*. Ph.D. Dissertation, University of Edinburgh. College of Science and Engineering. School of Informatics.

Dagan, I.; Lee, L.; and Pereira, F. 1997. Similarity-based methods for word sense disambiguation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, 56–63. Association for Computational Linguistics.

Greene, S., and Resnik, P. 2009. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 503–511.

Grefenstette, G. 1994. *Explorations in automatic thesaurus discovery*. Springer.

Gupta, D. 2008. *Understanding Terrorism and Political Violence: The Life Cycle of Birth, Growth, Transformation, and Demise*. New York: Routledge.

Joachims, T. 2002. *Learning to Classify Text Using Support Vector Machines*. Norwell, MA: Kluwer.

Kahneman, D., and Tversky, A. 1979. Prospect theory: An analysis of decision under risk. *Econometrica* XLVII:263–91.

Kahneman, D., P. S., and Tversky, A. 1982. *Judgment Under Uncertainty: Heuristics and Biases*. New York: Cambridge University Press.

Koppel, M., and Schler, J. 2006. The importance of neutral examples for learning sentiment. *Computational Intelligence* 22(2):100–109.

Lewis, D. D. 1992. An evaluation of phrasal and clustered representations on a text catego- rization task. In *Proceedings of SIGIR-92*, 37–50.

Lin, D. 1998. Automatic Retrieval and Clustering of Similar Words. In *Annual Meeting-Association for Computational Linguistics*, volume 36, 768–774. Association for Computational Linguistics.

Malouf, R., and Mullen, T. 2007. Graph-based user classification for informal online political discourse. In *Proceedings of the 1st Workshop on Information Credibility on the Web (WICOW)*.

Manning, C.; Raghavan, P.; and Schütze, H. 2008. *An Introduction to Information Retrieval* . Cambridge, UK: Cambridge University Press.

Mullen, T., and Collier, N. 2004. Sentiment analysis using support vector machines with diverse information sources. In *Proceedings of EMNLP*, volume 2004, 412–418.

Nivre, J. 2003. An Efficient Algorithm for Projective Dependency Parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT 03)*, 149–160.

Paltoglou, G., and Thelwall, M. 2010. A study of information retrieval weighting schemes for sentiment analysis. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1386–1395. Association for Computational Linguistics.

Pang, B., and Lee, L. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, 115–124. Association for Computational Linguistics.

Pantel, A. P. 2003. *Clustering by Committee*. Ph.D. Dissertation, University of Alberta.

Prabowo, R., and Thelwall, M. 2009. Sentiment analysis: A combined approach. *Journal of Infometrics* 3(1):143–157.

Sageman, M. 2004. *Understanding Terror Networks*. Philadelphia: University of Pennsylvania Press.

Salton, G., and McGill, M. 1983. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Schütze, H. 1993. Part-of-speech induction from scratch. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, 251–258. Association for Computational Linguistics.

Silver, N., and Dunlap, W. 1987. Averaging correlation coefficients: Should fisher's z transformation be used? *Journal of Applied Psychology* 72(1):146–148.

Tajfel, H. 1978. *Differentiation between Social Groups: Studies in Inter-Group Relationship*. London: Academic Press.

Tajfel, H. 1981. *Human Groups and Social Categories: Studies in Social Psychology*. Cambridge: Cambridge University Press.

Tsou, M.-H. 2011. Mapping cyberspace: Tracking the spread of ideas on the internet. In *Proceeding of the 25th International Cartographic conference*. International Cartographic Association.

Turney, P., and Pantel, P. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37:141–188.

Turney, P. 2002. Thumbs up or thumbs down? semantic orientation applied to unnsupervised classification of reviews. In *Proceedings of the ACL-02*, 417–424.

Whitelaw, C.; Garg, N.; and Argamon, S. 2005. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, 625–631. ACM.

Zhitomirsky-Geffet, M., and Dagan, I. 2009. Bootstrapping distributional feature vector quality. *Computational Linguistics* 35(3):435–461.

Zhou, Y.; Qin, J.; Lai, G.; and Chen, H. 2007. Collection of us extremist online forums: A web mining approach. In *40th Annual Hawaii International Conference on System Sciences*, 70–70. IEEE.