Identifying Political Sentiment between Nation States with Social Media

Nathanael Chambers, Victor Bowen, Ethan Genco, Xisen Tian,
Eric Young, Ganesh Harihara, Eugene Yang
Department of Computer Science
United States Naval Academy
nchamber@usna.edu

Abstract

This paper describes an approach to large-scale modeling of sentiment analysis for the social sciences. The goal is to model relations between nation states through social media. Many cross-disciplinary applications of NLP involve making predictions (such as predicting political elections), but this paper instead focuses on a model that is applicable to broader analysis. Do citizens express opinions in line with their home country's formal relations? When opinions diverge over time, what is the cause and can social media serve to detect these changes? We describe several learning algorithms to study how the populace of a country discusses foreign nations on Twitter, ranging from state-of-the-art contextual sentiment analysis to some required practical learners that filter irrelevant tweets. We evaluate on standard sentiment evaluations, but we also show strong correlations with two public opinion polls and current international alliance relationships. We conclude with some political science use cases.

1 Introduction

The volume of text available on social media provides a new opportunity for public policy and political science. Specifically in the area of international relations, advances in natural language understanding and sentiment analysis may offer new insights into the sentiment of one nation toward another. This paper processes 17 months of Twitter data to identify discussions about sovereign states, and it aggregates opinions toward these states from foreign nations. We present a novel application of contextual sentiment with this task, and identify several semi-supervised learning algorithms that are needed to address the reference resolution challenge inherent to country names. We present intrinsic evaluations of our learners on labeled datasets as well as four extrinsic political science evaluations that show strong alignment with our large-scale sentiment extraction.

An open question for international policy makers is the extent to which public opinion drives decision making. How do military conflicts affect a neutral nation's relationship? Does public opinion shift toward a country after a formal alliance is created, or must popular opinion shift first? These questions are difficult to address due to the lack of measurable data. While polling data can be collected, collection beyond a handful of countries is cost prohibitive. This paper hypothesizes that sentiment analysis can be used as a proxy to track international relations between nation states. We describe the largest attempt (over 2 billion tweets) to measure nation state sentiment across hundreds of country pairs.

The core challenge to measuring public opinion between countries is an accurate algorithm to judge the sentiment of a text toward another nation. Unlike traditional sentiment analysis, the general sentiment of the text is not adequate. Let the following serve as an example.

> *I miss **Pakistan**. I am in full sad mode right about now.* @RachOrange (California)

This tweet is a *positive* example from the USA toward Pakistan. However, a typical sentiment classifier misclassifies this as negative because *miss* and *sad* express sadness. A *contextual* sentiment classification is needed to identify that the predicate *miss* is positive toward its argument. Several

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 65–75,
Lisbon, Portugal, 17-21 September 2015.

recent competitions included contextual classification tasks, and this paper builds on the best of those algorithms for a unique nation-nation sentiment classifier. We describe a multi-classifier model that aggregates tweets into counts of positive and negative sentiment from one country toward another. Several unique filters are required to resolve textual references toward country names.

We first present standard NLP sentiment experiments that show the classifiers achieve good performance on individual tweets. To evaluate the complete nation-nation system, we present four novel evaluations, including two public opinion polls. Correlation with the polls is high at $\rho = .8$, and our nation-nation sentiment is 84% accurate with NATO and EU relations. We then discuss the implications for both NLP as a technical science and political science as a social science.

2 Previous Work

Sentiment analysis is a large field applicable to many genres. This paper focuses on social media and contextual polarity, so we only address the closest work in those areas. For a broader perspective, several survey papers are available (Pang and Lee, 2008; Tang et al., 2009; Liu and Zhang, 2012; Tsytsarau and Palpanas, 2012).

Several sources for microblogs have been used to measure a large population's mood and opinion. O'Connor et al. (2010) used Twitter data to compute a ratio of positive and negative words to measure consumer confidence and presidential approval. Kramer (2010) counted lexicon words on Facebook for a general 'happiness' measure, and Thelwall (2011) built a general sentiment model on MySpace user comments. These are early general sentiment algorithms for social media.

Other microblog research focused on finding noisy training data with distant supervision. Many of these algorithms use emoticons as semantic indicators of polarity. For instance, a tweet that contains a sad face likely contains a negative polarity (Read, 2005; Go et al., 2009; Bifet and Frank, 2010; Pak and Paroubek, 2010; Davidov et al., 2010; Kouloumpis et al., 2011). In a similar vein, hashtags can serve as noisy labels (Davidov et al., 2010; Kouloumpis et al., 2011). Our bootstrap learner is similar in its selection of seed tokens.

Supervised learning for *contextual* polarity has received more attention recently. Jiang et al. (2011) is an early approach. Work on product reviews sought the sentiment toward particular product features. These systems used rule based models based on parts of speech and surface features (Nasukawa and Yi, 2003; Hu and Liu, 2004; Ding and Liu, 2007). Most notably, recent Semeval competitions addressed contextual polarity (Nakov et al., 2013; Rosenthal et al., 2014). The top performing systems learned their own lexicons custom to the domain (Mohammad et al., 2013; Zhu et al., 2014). Our proposed system includes many of their features, but several fail to help on nation-nation sentiment.

Early approaches to *topic detection* on social media were straightforward, selecting a keyword (e.g., "Obama") to represent the topic (e.g., "US President") and retrieving tweets containing the word (O'Connor et al., 2010; Tumasjan et al., 2010; Tan et al., 2011). These systems classify the polarity of *the entire tweet*, but ignore the question of polarity toward the particular topic. This paper focuses on identifying tweets with nation mentions, and identifying the sentiment toward the mention, not the overall sentiment of the text.

Event detection on Twitter is also relevant (Sakaki et al., 2010; Becker et al., 2011). In fact, O'Connor et al. (2013) modeled events to detect international relations, but our goal is to model long term relation trends, not isolated events.

Large-scale computational studies of social media are relatively new to the international relations community. Barbera and Rivero (2014) is a notable example for election analysis. Some studied online discussion about Palestine (Lynch, 2014) and the role of Twitter in the Arab Spring (Howard et al., 2011; Howard, 2013). However, they simply counted the volume of tweets containing keywords. This paper applies a deeper NLP analysis and we show that frequency alone fails at detecting nation-nation relations.

Most relevant to this paper is a study of Arabic tweets into anti-American sentiment. Jamal et al. (2015) used a supervised sentiment classifier on Arabic tweets to measure sentiment toward the USA. Our paper differs by taking a broader view. We investigate with state-of-the-art sentiment algorithms, and we we study practical problems that arise within when measuring nation-nation sentiment across *all* country pairs. To our knowledge, this paper is the largest computational approach (17 months with 2 billion tweets) to measuring international relations on social media.

3 Microblog Datasets

The main dataset for this study is 17 months of tweets obtained through the keyword Twitter API that mention one of 187 unique countries. The dataset spans from Sep. 3, 2013 to Jan 10, 2015 with 3-5 million tweets per day. Each tweet includes the profile location and geolocation data (if available) of the user who posted the tweet. Collection was not limited to a specific location in order to retrieve samples from across the world. This dataset is used in all political science experiments (Sections 6.2 and 6.3).

A smaller labeled dataset is used for supervised classification. We randomly sampled the data to create a dataset of 4250 tweets. The authors initially labeled each tweet with one of four sentiment labels: *positive*, *negative*, *objective*, or *irrelevant*. Text was only labeled as positive if it is positive toward the nation's mention. Text that contains a nation's mention, but does not contain sentiment toward the mention is labeled *objective*. Text with a mention that is *not* referent to a physical country is labeled *irrelevant* despite presence of sentiment. This irrelevant distinction is a departure from sentiment competitions. A second labeling added a fifth label to the *irrelevant* tweets to split off *dining* topics.

Usernames (e.g., @*user*) and URLs are replaced with placeholder tokens. Multiple whitespace characters are condensed and the text is split on it. Punctuation attached to tokens is removed (but saved) and used in later punctuation features. Punctuation is not treated as separate tokens in the n-gram features. We prepend occurrences of "not" to their subsequent tokens, merging the two into a new token (e.g., "not happy" becomes "not-happy"). Once the raw text of the tweet is tokenized as above, non-English tweets are filtered out. English filtering is performed by LingPipe[1]. We manually evaluated this filter and found it 86.2% accurate over 400 tweets. Accuracy is lost due to slang and the short nature of the text.

4 Classifying Nation-Nation Sentiment

Given a tweet containing a country's name, our goal is to identify the sentiment of the text toward that nation. Unlike most work on contextual polarity, this requires reference resolution of the target phrase (e.g., the country name). Previous Semeval competitions evaluate the sentiment of a text toward a phrase, but the semantics of the phrase is largely ignored. For instance, the following example would make an excellent Semeval test item, but its classification is irrelevant to the goal of measuring nation sentiment:

> *My daughter and I have been to Angelo's several times when in Little **Italy**. Love love it!*

The author is obviously positively inclined toward Little Italy, however, Little Italy does not refer to the country of Italy. We found that most tweets referring to dining or visiting foreign-themed restaurants are not relevant to determining nation to nation sentiment. It became necessary to research new classifiers that perform basic reference resolution.

4.1 Reference Resolution: Irrelevant Detection

This paper defines reference resolution in the traditional linguistic sense: determine the real-world referent of a text mention. Most NLP tasks use coreference resolution: determine the text antecedent of a text mention. This paper requires reference resolution because the target phrase often does not refer to an actual geolocated country. After collecting months of tweets that include country name mentions, data analysis revealed several types of these non-references. We treat reference resolution as a classification problem. Below are a variety of supervised and semi-supervised learners that identify different types of errant country references, and ultimately serve to filter out these irrelevant tweets.

4.1.1 Dining Classifier

One of our early observations was that mentions of nations are often in the context of eating and dining, as evidenced here:

> *This is the first **turkey** sandwich I've had in awhile... It's great **turkey**.*

> *Taste of **China** For chinese food lover's. For more info Please visit*

This class of tweet is problematic to our study of international politics. While microblogs about dining can contain heavy emotion, a link to the writer's opinion about the foreign nation itself is ambiguous. We thus filter out dining text through supervised classification. Using the labeled dataset in Section 3, we annotated a *dine* label for all dining tweets. Tweets without a *dine*

[1] *alias-i.com/lingpipe/#lingpipe*

	Dine	Rel
All unigrams in text	✓	✓
1-3grams that include the country	✓	✓
Bigram and Trigram country pattern	✓	✓
Four Square app pattern		✓
Named Entity 2-3grams w/ country		✓
Emoticon happy or sad		✓
Ending text punctuation		✓
Binary: contains exclamation point		✓

Table 1: Dining and Relevant features.

label are considered *not-dine*. We ran a logistic regression for two labels, *dine* and *not-dine*. Text features are shown in Table 1.

4.1.2 Irrelevancy Classifier

Beyond dining, a broader class of irrelevant tweets refer to non-nation entities. These microblogs contain country names, but the mentions do not reference the physical country. The following examples illustrate this class of *irrelevant* tweets (nation tokens in bold):

> *Yesterday was chilly out and now today's going to be 80. New **England** weather is so bipolar I hate it so much*

> *Bank Of **America** Upgrades ConocoPhillips On More Favorable Outlook*

Several types of irrelevancy can be found, but the most common is a non-nation geolocation like *New England*. Proper nouns like *Bank of America* are frequent as well. A named entity recognizer (NER) identified some of these, but we ultimately turned to supervised classification for better accuracy (space constraints prevent discussion of NER performance). We trained a logistic regression classifier on the **relevant** tweets in the Section 3 dataset, and mapped all other labels to **irrelevant**. Features used are shown in Table 1.

4.1.3 Bootstrap Learner

After filtering non-referent tweets, we observed that many *positive* and *negative* tweets reference countries in the context of sporting events and music/concerts. These are correctly labeled **relevant** by the above binary classifiers (and possibly annotated as positive or negative), but the topic (sports or music) does not contain a strong semantic connection to the author's actual opinion about the country. A couple of sport examples are given here:

> *@SpecialKBrook Wow - the Brittish judge scored the fight a draw - lucky **England**'s fighters are better than their judges.*

> ***Congo** LFC now someone give me that goalie's jersey :p*

The sport topic has a less diverse vocabulary than other topics. We hypothesized that a bootstrap learning framework (Riloff and Jones, 1999) could quickly learn its unique language without the need for supervised learning. Beginning with a short list of sport keywords (*football, basketball, baseball, cricket, soccer, golf, hockey, rugby, game, vs*), we ran two iterations of a bootstrapped learner. The first step retrieves tweets containing one of the keywords. The second counts token occurrences in this set and computes pointwise mutual information (PMI) scores for each unigram by comparing with the unigram counts over the entire corpus. The learner processed ˜190 million tweets (a couple months of data). The PMI scores from this process then form the basis of a simple topic classifier.

A tweet is classified as a topic (e.g., sports) if its average token PMI score is above a learned threshold for that topic:

$$score_T(text) = \frac{1}{N} \sum_{w \in text} pmi_T(w) \qquad (1)$$

where N is the number of tokens in the text and $T \in \{sports, concerts\}$. The text is classified as in topic if $score_T(text) > \lambda_t$. The threshold λ_T was determined by visual inspection of a held out 1000 tweets to maximize accuracy. The initial seed words and λ_T thresholds for each topic are given here:

Seed Words	λ
football, basketball, baseball, cricket, soccer, golf, hockey, rugby, game, vs	0.08
concert, music, album, song, playlist, stage, drum	0.15

4.2 Contextual Sentiment Analysis

The above classifiers identify *relevant* tweets with *references* to geolocated nations. Approximately 21% are filtered out, leaving 79% for the remaining component of this paper: contextual sentiment analysis. Contextual sentiment analysis focuses on the disposition of text toward a word or phrase (in this case, a country's name). Most data-driven approaches rely on labeled corpora to drive the learning process, and this paper is no different.

Assigning polarity to a word/phrase requires features that capture the surrounding context. The following tweets are examples of context with strong polarity toward the country in bold.

*RT @ChrissyCostanza: Happiest girl ever. I LOVE YOU **SINGAPORE***

*there's no **Singapore** Got Talent cus the only talent we have is stomping complaining & staring*

Singapore is the target country here. The first tweet is overtly positive toward it, but the second requires a more subtle interpretation. The negative context is toward *us*, referencing the people of the Singapore anaphor. It seems reasonable to infer that they are negative toward the country as a whole, but a deeper reasoning is required to make the connection. These difficult decisions require a wide-range of lexical features. We build on the top performing features from contextual polarity systems in Semeval 2013 and 2014 (Mohammad et al., 2013; Zhu et al., 2014). We used the following set of features to capture these different contexts:

Token Features: All unigrams and bigrams.

Target Patterns: This feature creates patterns from n-grams that include the target word. The target is replaced with a variable to capture generalized patterns. For instance, "to France last" becomes "to X last". Bigram and trigram patterns are created.

Punctuation: End of sentence punctuation and punctuation attached to target words. Prefix and postfix punctuation are separate features.

Emoticons: Two binary features for the presence/absence of smiley and sad face emoticons.

Hand-Built Dictionary: Two binary features, *postivemood* and *negativemood*, indicate if a token appears in a sentiment lexicon's positive or negative list. We use Bing Liu's Opinion Lexicon[2].

Nation-Nation Learned Dictionary: Following the success of Zhu et al. (2014), we learn a mood dictionary from our domain-specific nation dataset. We count unigrams (bigrams did not improve performance) in one year of unfiltered tweets with nation mentions that contain an emoticon. Using these counts, each unigram computes its PMI scores toward happy and sad contexts. We construct features based on these PMI scores: (1) the highest happy PMI score of all unigrams in a tweet, (2) the highest sad PMI score, (3) the number of positive tokens, (4) the number of negative tokens, and (5) the sum of the token PMI differences between happy-sad.

[2] *http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html*

General Learned Dictionary: We computed the same features as in the above learned dictionary, but instead counted tokens in all tweets of the general emoticon corpus of Go et al. (2009).

The contextual sentiment learner is trained on the labeled dataset (Section 3). Only tweets with *positive*, *negative*, or *objective* labels are included (irrelevant and dining are ignored). Stanford's CoreNLP (nlp.stanford.edu/software) is used to train a MaxEnt classifier with its default settings.

5 Nation to Nation Pipeline

The complete system to determine the nation-nation sentiment of a tweet consists of 3 steps: (1) identify the country origin of the tweet, (2) filter out tweets without references to geolocated nations and filter out irrelevant topics, and (3) identify the sentiment toward the country. We processed 17 months of tweets (Section 3).

The first step identifies the origin of the tweet with either its GPS coordinates or the profile location of the Twitter user. Profile locations are mapped to countries with an exhaustive list of country names, major cities, and patterns that match US city/states (e.g., *Pensacola,FL* maps to *USA*). Non-english tweets are removed with lingpipe. The second step filters non-referent, irrelevant, dining, and concert tweets with the classifiers from Section 4.1 (about 21% of tweets at this stage). The final step is the contextual sentiment classifier (Section 4.2). Tweets that make it through receive 1 of 3 possible labels: positive, negative, objective.

The aggregate counts of the three labels are collected for each day. This was around 1.2 million nation labels per day over 17 months. The counts are used for evaluation in the experiments.

6 Experiments

Our goal is to first prove the accuracy of our sentiment classifiers, then show the broader pipeline's correlation with known nation-nation politics.We thus conducted three types of experiments. The first is an intrinsic evaluation of the classifiers with common frameworks from the NLP community. The second is an extrinsic evaluation from multiple political science datasets. The third is a set of use case proposals for application of this analysis.

Dining Classifier

Label	P	R	F1
dining	.76	.48	.59
not-dining	.96	.99	.98
Baseline Accuracy		93.1%	
Accuracy		95.3%	

Irrelevant Classifier

Label	Prec	Recall	F1
irrelevant	.84	.90	.87
relevant	.84	.75	.80
Baseline Accuracy		58.7%	
Accuracy		84.0%	

Sentiment Classifier

Label	Prec	Recall	F1
positive	.64	.47	.54
negative	.60	.33	.43
objective	.71	.87	.78
Baseline Accuracy		59.0%	
Accuracy		68.7%	

Table 2: Classifier performance. Precision/Recall is calculated for each label separately. Accuracy is over all labels: # correct/total.

6.1 Classifier Experiments

The dining, irrelevant, and sentiment classifiers are supervised systems trained on a labeled dataset of 4,250 tweets. We split the dataset into training, dev, and test sets. The dev set contains 200 tweets, the test set has 750 tweets, and the training set size varied based on the available labels. The features in this paper were developed solely on the training and dev datasets. Reported results are on the unseen test set of 750 tweets. The bootstrapped classifiers for sports and concerts were learned without labeled data, so we ran the sports and concerts classifiers on an unseen portion of our data, and manually evaluated the first 200 tweets that were labeled by each classifier.

Precision and recall are calculated individually for each class label: $P = \#correct/\#guessed$ and $R = \#correct/\#gold$. Where $\#guessed$ is how many times the classifier predicted the target label, and $\#gold$ is how many times the target label appears in the dataset. Accuracy is also shown, calculated as a single score over all labels together: $Accuracy = \#correct/N$. The first table in Table 2 shows the dining classifier's performance. The majority class baseline is high at 93% because only 7% of the data is about dining. The classifier achieves a 29% decrease in accuracy error (2% absolute increase). The second table shows the more general irrelevant classifier. The majority class baseline is much lower than dining at 58.7%. Many tweets that contain a country name are not relevant nor references to the geolocated country itself. Our trained classifier does well on this task achieving 84% accuracy, a 26% absolute increase over baseline. It is 84% precise with 90% recall on detecting irrelevant tweets. The third table in Table 2 shows sentiment classifier results. Accuracy is almost 10% absolute above the majority class.

Finally, the bootstrapped classifiers perform at 98% accuracy for sports and 90% for concerts.

Positive/Negative Ratios

Target	Ratio	Target	Ratio
US to Canada	11.9	US to Ireland	3.2
US to Italy	10.8	US to Spain	3.0
US to Japan	7.7	US to France	2.7
US to Australia	3.5	US to Jordan	2.1
US to UK	3.5	US to Mexico	1.9

Table 3: Positive/Negative ratios for the US toward its top 10 frequently mentioned nations.

6.2 Nation-Nation Sentiment Experiments

Nation opinions are represented as directed edges: each edge (X,Y) represents the opinion of nation X toward nation Y. The weight of an edge is the ratio of positive to negative counts:

$$R(X,Y) = C(X,Y,positive)/C(X,Y,negative)$$

where C(X,Y,L) is the number of tweets by nation X users about nation Y with sentiment L. Only tweets that make it through the Nation to Nation Pipeline of Section 5 receive sentiment labels. If a nation pair (X,Y) was observed less than 1000 times, it is not included in the evaluations. We provide experiments later to evaluate this cutoff's affect.

The dataset (Section 3) spans 17 months from 2013-2015. All tweets are classified or filtered out, and R(X,Y) is computed for all pairs. Table 3 shows the top 10 nation pair ratios (with over 500k tweets between them) for the U.S.

We present four formal evaluations to answer the central question of this paper: can sentiment from social media be used to help approximate international opinions? The first two experiments use public opinion polls of national sentiment toward other nations. The third uses military conflicts as a proxy for *negative* relations, and the fourth uses current formal alliances as a proxy for *positive* relations. None of these can provide a complete picture of the connection between popular sentiment and international relations, but the four together provide a strong case that sentiment contains a useful signal.

Correlation: Public Opinion Polls

Human Poll	Sentiment	Freq. Baseline
Germany	Canada	China
Canada	Japan	Israel
UK	EU	USA
Japan	France	Russia
France	UK	India
EU	Brazil	Japan
Brazil	USA	Canada
USA	India	UK
China	South Africa	Pakistan
South Korea	South Korea	France
South Africa	Germany	Iran
India	Russia	Brazil
Russia	China	Germany
Israel	Israel	North Korea
North Korea	North Korea	South Korea
Pakistan	Iran	South Africa
Iran	Pakistan	EU
Correlation	**0.80**	**-0.06**

Table 4: Polling Data: ranking of a nation's "positive contribution" to the world, compared to automatically identified nation-nation sentiment.

Each year, GlobeScan/PIPA releases polling data of 16 nations in a ranked ordering based on how 26,000 people view their "positive contribution" to the world[3]. This poll helps to determine whether or not this paper's sentiment pipeline matches human polling. We created our own ranking by assigning a world score to each nation n: the average sentiment ratio of all other nations toward n. Since the polling data also ranks the EU, we average the EU member nation world scores for an EU world score. Table 4 shows the PIPA poll (**Human Poll**) and our world ranking (**Sentiment**). Using Spearman's rank correlation coefficient to measure agreement, our ranking is strongly correlated at $\rho = 0.8$ (perfect is 1.0). The main mistake in our ranking is Germany. We also compare against a **Frequency Baseline** to eliminate the possibility that it's simply a matter of topic popularity. Poll rankings could simply be correlated with who people choose to discuss, or vice versa. The frequency baseline is the average number of twitter mentions per nation (i.e., the most discussed). This baseline shows no correlation at $\rho = -.06$.

We then evaluated against a US-centric polling agency, Gallup. They asked Americans to rate other nations as 'favorable' or 'unfavorable' in a 2014 poll[4]. The result is a ranking of favorability. In contrast to the PIPA poll which evaluates many nations looking in, Gallup evaluates a single nation looking out. Space constraints prevent us from visually showing the US ranking, but again the sentiment ratios have a strong correlation at $\rho = .81$. The frequency baseline is $\rho = .23$. The nation-nation sentiment extraction strongly correlates with both world views (PIPA) and US-specific views (Gallup).

The third evaluation uses a political science dataset from the Correlates of War Project, the **Militarized Interstate Disputes v4.01** (MID) (Ghosn et al., 2004). This dataset is used in the field of international relations, listing conflicts since 1816. We limit the evaluation to conflicts after 1990 to keep relations current. The dataset ends at 2001, so while not a completely current evaluation, it stands as a proxy for negative relations. Each conflict in MID is labeled with a conflict severity. We convert severity labels between nations to a pair score MID(X,Y):

$$MID(X,Y) = \sum_{d \in Disputes(X,Y)} score(d) \quad (2)$$

where Disputes(X,Y) is the set of conflicts between the two nations X and Y, and score(d) is a severity score for the type of dispute d. *War* is -5, *use of force* is -4, *displays of force* is -3, *threatening use of force* is -2, and *no militarized action* is -1. We take the sum of severity scores and save all nation pairs (X,Y) such that $MID(X,Y) < -10$. This score indicates multiple conflicts and are thus considered as nations with true negative relations.

We then compare our sentiment ratios R(X,Y) against these gold negative pairs. Each continuous R(X,Y) is discretized into sentiment categories for ease of comparison. Since the mean across all R(X,Y) is 1.25, we consider an interval around 1.25 as neutral and create positive and negative labels above and below that neutral center:

$$ratiolabel(Z) = \begin{cases} positive, & \text{if } Z \geq 2.4 \\ slightpos, & \text{if } 2.4 > Z \geq 1.4 \\ neutral, & \text{if } 1.4 > Z \geq 1.1 \\ slightneg, & \text{if } 1.1 > Z \geq 0.8 \\ negative, & \text{if } 0.8 > Z \end{cases}$$

The bottom table in Table 5 shows the number of nation pairs that align with the negative labels of the MID dataset. Only pairs that have at least 1000 tweets are evaluated. Of the resulting 90 pairs, 61 are correctly identified by our system as negative or slight negative for an **accuracy of 68%**. 19 *positive* pairs are incorrectly aligned with *MID-negative*. Error analysis shows that many incorrect

[3] http://www.worldpublicopinion.org/pipa/2013CountryRatingPoll.pdf
[4] http://www.gallup.com/poll/1624/perceptions-foreign-countries.aspx

Positive: Formal Alliances

	Pos	SlPos	N	SlNeg	Neg
# Nation Pairs	**341**	**65**	22	**26**	**28**

Negative: Military Disputes

	Pos	SlPos	N	SlNeg	Neg
MID-Negative	**12**	**7**	10	**15**	**46**

Table 5: Top: The number of NATO/EU nation pairs with automatic sentiment labels. Bottom: The number of pairs with military disputes (MID dataset) and automatic sentiment labels.

labels are between nations with a smaller Twitter presence, so performance likely suffers due to lack of data. For robustness testing, we shifted the thresholds that discretize the nation ratios and MID scores into postive and negative categories. The accuracy result shows little change. We also reran the experiment with a higher cutoff of 10,000 instead of 1,000. The negative disputes **accuracy increases from 68% to 81%**, but the recall obviously drops as less countries are included. This suggests the sentiment ratios might be used on a sliding confidence scale based on frequency of mention.

To evaluate positive relations, we use current alliances as a fourth evaluation. NATO and the EU are the main global alliances with elements of mutual defense. We do not include trade-only alliances as trade is not always an indication of allegiance and approval (Russia and Ukraine is a current example of this disparity). This evaluation considers pairs of nations within NATO and within the EU as gold positive relations. We compare our sentiment ratios to these pairs in the top of Table 5. This evaluation is broader than the conflict evaluation because NATO and EU nations have more of a Twitter presence. Of the 482 country pairs, our positive/slightpos accuracy is **84.2%**.

6.3 Application Experiments

We now briefly discuss how these positive results for nation-nation sentiment relates to political science analysis.

One core area of study is how national sentiment shifts over time, *and why*. Computing R(X,Y) on a bi-weekly basis, Figure 1 graphs the sentiment ratio from the USA toward India and Israel. The timeline shows significant favorability toward India during their extended election season, but afterward the opinion is similar to before the election. In contrast, the 2014 Israel-Gaza conflict shows a very different effect. US opinion of Israel is initially steady (slightly positive) until the conflict causes a large dip. Unlike India's spike, *US opinion stays depressed* even after the conflict concludes. It appears to have only risen to 'normal' levels months later. We do note that the water is slightly muddied because our algorithm may not distinguish well between sentiment toward the war, Israel itself, or even sympathy toward casualties. However, it's clear that nation-nation sentiment is captured, and future work is needed to identify finer grained sentiment as needed.

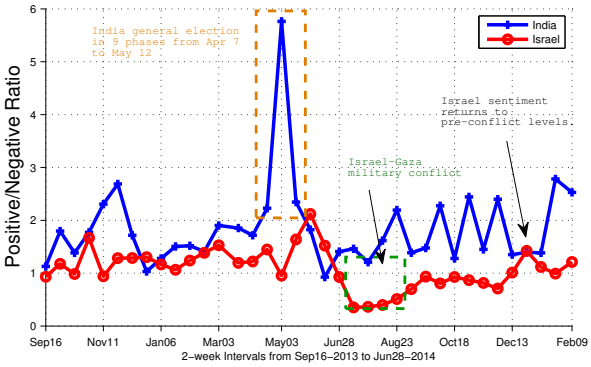

Figure 1: USA opinion of India/Israel over 2-week intervals from Sep-2013 to Feb-2015.

Another application is inter-alliance relations. For instance, Table 6 shows how NATO member nations view *other* alliances. The table shows the average of all R(X,Y) edges for each nation within an alliance to a nation in the other. According to our ratios, NATO countries have stark differences between how they view themselves versus how they view African Union/Arab League nations. Further, our pipeline enables analysis of outside nations looking in. For instance, the nations with the most positive view of the EU are Uruguay, Lithuania (EU member), Belarus, Moldova, and Slovakia (EU member). Almost all (not Uruguay) are eastern european nations. Moldova is currently seeking EU membership and Belarus had closer ties until recently. Our results might point to potential future alliances. Future work is needed to explore this implied connection.

Finally, the R(X,Y) ratios can also represent a nation's *opinion profile*. Represent each nation X by a vector of its R(X,Y) ratios. This represents its entire international view based on social me-

Inter-Alliance Opinion Ratios

Source	Target	Average R(X,Y)
NATO	African Union	0.45
NATO	Arab League	0.48
NATO	European Union	1.51
NATO	NATO	1.55

Table 6: Average pos/neg ratio of NATO nations toward the nations in other formal alliances.

MID Accuracy with Filters

Filters	Correct	Incorrect	Acc.
Dining+Sports	61	29	**68%**
Sports only	61	37	**62%**
None	56	51	**52%**

Table 7: Filtering effects on the MID results.

dia sentiment. Space prohibits more detail, but we clustered opinion profiles with k-means (k=12) and cosine similarity. Typical alliances, such as European and African clusters, are learned.

6.4 Ablation Tests

The sentiment pipeline includes two practical filters to remove tweets about **dining** and **sports**. We added these during training and developement solely based on our interpretation and analysis of the data. We did not evaluate on the test datasets until the very end. Table 7 shows results from the MID evaluation with the dining and sports filters removed in sequence.

The number of correctly identified negative nation pairs is mostly unchanged, but the number of incorrect decisions increases dramatically. This occurs because a greater number of tweets make it through the pipeline. Further, this shows that the filters effectively remove tweets that cause misclassification errors.

7 Discussion

This work is an important first step toward automatic means to broadly detect international relations from social media. We use sentiment analysis as a proxy for extracting at least one aspect of the large set of factors involved in such relations. This paper is the largest application of sentiment analysis across a diverse set of nation-nation pairs (hundreds of country pairs over 17 months), and we showed that this sentiment is strongly correlated ($\rho = 0.8$) with two independent public opinion polls. These correlations more importantly suggest that we are not simply identifying a binary positive or negative relation, but that the *relative sentiment scores* are useful. The failure of frequency baselines on this ranking further suggests that this is not a side effect of topic frequency.

One argument against using public opinion polls for an evaluation is that the same people who are polled by PIPA might be the same people who tend to voice opinions on Twitter. The Twitter dataset is not independent from the polls, so the strong correlation we found could simply be a matter of sampling the same population. This is not possible to know, but whether or not it is the case, this paper's pipeline could be quite valuable in automating expensive and time consuming human polls.

The results that focused on positive sentiment (polls and alliances) are quite high. Negative sentiment revealed a lower 68% accuracy on the MID dataset, but it is due to the fact that nation-nation conflicts often occur between smaller nations that are not represented well on Twitter. Requiring a higher observed count improves accuracy to 81%.

While we are cautious not to make broad claims about discovering international relations on Twitter, we are encouraged by the experimental alignment with current alliances and historical conflict data. The sentiment timeline for Israel and India (Figure 1) is also intriguing. Tracking nation relations over a longer time period presents an opportunity for future study. This continual tracking of sentiment is one of the most obvious benefits of an automated approach.

Finally, an interactive world map is available to browse this paper's data at *www.usna.edu/Users/cs/nchamber/nations*. Each nation can be selected to visually color the map with its positive/negative lens, and timelines showing sentiment shifts between nations are visible. All code, data, and results are also available on this page. We hope this work encourages even further connections between NLP and political science.

Acknowledgments

This work was supported in part by the DoD HPC Modernization Program. It also would not have been possible without the infrastructure support and help from the accommodating staff at the Maui HPCC. Special thanks to the reviewers and area chair for their parts in an exciting and fulfilling review process.

References

Pablo Barberá and Gonzalo Rivero. 2014. Understanding the political representativeness of twitter users. *Social Science Computer Review*, December.

Hila Becker, Mor Naaman, and Luis Gravano. 2011. Beyond trending topics: Real-world event identification on twitter. *ICWSM*, 11:438–441.

Albert Bifet and Eibe Frank. 2010. Sentiment knowledge discovery in twitter streaming data. In *Lecture Notes in Computer Science*, volume 6332, pages 1–15.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*.

Xiaowen Ding and Bing Liu. 2007. The utility of linguistic rules in opinion mining. In *Proceedings of SIGIR-2007*, pages 23–27.

Ghosn, Faten, Glenn Palmer, and Stuart Bremer. 2004. The mid3 data set, 1993-2001: Procedures, coding rules, and description. In *Conflict Management and Peace Science 21*, pages 133–154.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report.

Philip N. Howard, Aiden Duffy, Deen Freelon, Muzammil Hussain, Will Mari, and Marwa Mazaid. 2011. Opening closed regimes: What was the role of social media during the arab spring? Technical Report working paper 2011.1, University of Washington, September.

Philip N. Howard. 2013. *Democracy's Fourth Wave? Digital Media and the Arab Spring*. Oxford University Press.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Amaney A. Jamal, Robert O. Keohane, David Romney, and Dustin Tingley. 2015. Anti-americanism and anti-interventionism in arabic twitter discourses. *Perspectives on Politics*, pages 55–73, March.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the Association for Computational Linguistics (ACL-2011)*.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*.

Adam D. I. Kramer. 2010. An unobtrusive behavioral model of 'gross national happiness'. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI 2010)*.

Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. *Mining Text Data*, pages 415–463.

Marc Lynch. 2014. Arabs do care about gaza. www.washingtonpost.com/blogs/monkey-cage/wp/2014/07/14/arabs-do-care-about-gaza.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*.

Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. 2013. Semeval-2013 task 2: Sentiment analysis in twitter. In *Proceedings of the 7th International Workshop on Semantic Evaluation*. Association for Computational Linguistics.

Tetsuya Nasukawa and Jeonghee Yi. 2003. Sentiment analysis: capturing favorability using natural language processing. In *Proceedings of K-CAP*.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the AAAI Conference on Weblogs and Social Media*.

Brendan O'Connor, Brandon M Stewart, and Noah A Smith. 2013. Learning to extract international relations from political context. In *ACL (1)*, pages 1094–1104.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference On Language Resources and Evaluation (LREC)*.

B. Pang and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*.

Jonathon Read. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop (ACL-2005)*.

Ellen Riloff and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-99)*, pages 474–479.

Sara Rosenthal, Preslav Nakov, Alan Ritter, and Veselin Stoyanov. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. Association for Computational Linguistics.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM.

Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. 2011. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

H. Tang, S. Tan, and X. Cheng. 2009. A survey on sentiment detection of reviews. *Expert Systems with Applications*.

Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2011. Sentiment in twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418.

M. Tsytsarau and T. Palpanas. 2012. Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery Journal*, 24(3):478–514.

Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. Election forecasts with twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*.

Xiaodan Zhu, Svetlana Kiritchenko, and Saif M. Mohammad. 2014. Nrc-canada-2014: Recent improvements in the sentiment analysis of tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*.