Final Report: News Article Classifier

Christopher Walker
Eric Sun

June 10, 2015

Abstract

The goal of this project was to build a classifier that can distinguish articles relating to the Syrian Civil War as pro-regime or anti-regime based on their contents. Bigram and unigram feature sets were both tested. Naive Bayes, Maximum Entropy, Decision Tree, and SVM methods were tested. Repeated Random Sub-Sampling was used for validation. The results indicated no significant difference between unigram and bigram analysis. Naive Bayes is shown to have the highest performance for low quantity of training data, while SVM is shown to have the highest performance for high quantity of training data. A set of vocabulary more common among anti-regime media than pro-regime media is found.

Overview

Our task was to predict whether a news article is favorable to or opposed to the Syrian regime based on the vocabulary used by the article. The task is important as it can help determine the biases inherent in media coverage of the Syrian civil war by both pro- and anti-regime news sources. It is hoped that this can help determine whether there are specific sets of vocabulary used by pro- and anti-regime media.

Input Data

Data used to date includes mostly pro-regime articles from RT, Xinhua, and Global Times, as well as mostly anti-regime articles from BBC, Guardian, Washington Post, Fox, and CNN. A total of 703 articles was examined, including 216 pro-regime articles and 487 anti-regime articles. The chosen articles were restricted to those directly relating to the Syrian civil war or unrest in Syria. Only articles from early 2012 to late 2013 were chosen, as articles since 2013 have tended to focus on the actions of ISIL rather than the actions of the regime. Only English language articles were used.

In this project, Xinhua, RT, and Global Times are generally considered pro-regime despite the fact that Xinhua and Global Times may be considered more neutral in comparison to Russian media.

Attributes

Attribute sets were based on unigrams and bigrams. The attributes were restricted to the 500 most frequent attributes. To distinguish between common, irrelevant attributes (”the”) and common, relevant attributes (”militants”), a list of approximately 100 stopwords was generated and tested for the unigram analysis. No stopwords were used in the bigram analysis.

Methods and Validation

Five different methods were examined. Nearest Neighbor was tried early on but produced poor results, so it was abandoned. The final data were collected using 4 different methods, including Naive Bayes, Multinomial Logistic Regression (Maximum Entropy), Decision Tree, and SVM. The Nearest Neighbor was implemented manually, whereas Naive Bayes, Maximum Entropy, and Decision Tree methods were included in the NLTK toolkit in Python. The SVM method was included in the Sklearn toolkit in Python.

The training and validation method used was Repeated Random Sub-Sampling Validation. For each method, 9 different training-testing data splits were tried, with 10 trials for each split. The average accuracy across all 10 trials is the accuracy that is recorded. This validation method was chosen because of both the simplicity of implementing it as well as the flexibility of this method in handling data sets where there is not a 50-50 split in the classes present.

Results

Using stopwords in the unigram analysis decreased accuracy of the classifier for Naive Bayes, Decision Tree, and SVM methods, but had no significant effect for the Maximum Entropy method. The stopword list included unigrams such as ”since” or ”particularly” which were not relevant to the Syrian Civil War but which were nevertheless helpful to the classifier (for example, ”particularly” was far more common in American and British pro-regime sources than in Russian RT or Chinese Xinhua or Global Times.

Using bigram attributes without stopwords did not significantly change accuracy for any of the classifiers in comparison with unigram attributes without stopwords.

The ratio of anti-regime articles to all articles used was 0.693, so ZeroR accuracy was about 0.69 in most trials. Naive Bayes consistently outperformed ZeroR, achieving accuracy near 0.8 in most cases. Maximum Entropy performed poorly, and regularly matched ZeroR in performance. Decision Tree had poor performance with lower training ratios (ratio of data used for training to all data used) but matched performance of Naive Bayes for higher training ratios. SVM had performance similar to ZeroR for the lowest training ratio but had the best performance for higher training ratios (approaching 0.9 accuracy).

Accuracy Data

Table 1: Accuracy with unigrams, using stopwords

Training Ratio	Naive Bayes	Maximum Entropy	Decision Tree	SVM
0.1	0.791	0.699	0.704	0.703
0.2	0.792	0.696	0.730	0.742
0.3	0.798	0.696	0.733	0.775
0.4	0.799	0.695	0.769	0.811
0.5	0.799	0.693	0.758	0.829
0.6	0.798	0.691	0.763	0.831
0.7	0.805	0.705	0.785	0.847
0.8	0.801	0.689	0.783	0.852
0.9	0.815	0.686	0.785	0.867

Table 2: Accuracy with unigrams, not using stopwords

Training Ratio	Naive Bayes	Maximum Entropy	Decision Tree	SVM
0.1	0.813	0.705	0.739	0.709
0.2	0.814	0.699	0.763	0.778
0.3	0.820	0.698	0.803	0.846
0.4	0.813	0.691	0.792	0.860
0.5	0.806	0.708	0.808	0.882
0.6	0.814	0.697	0.811	0.897
0.7	0.806	0.693	0.809	0.890
0.8	0.829	0.690	0.823	0.912
0.9	0.806	0.703	0.831	0.923

Table 3: Accuracy with bigrams, not using stopwords

Training Ratio	Naive Bayes	Maximum Entropy	Decision Tree	SVM
0.1	0.800	0.692	0.741	0.692
0.2	0.788	0.688	0.796	0.688
0.3	0.786	0.688	0.796	0.717
0.4	0.784	0.692	0.816	0.771
0.5	0.783	0.698	0.809	0.838
0.6	0.787	0.693	0.826	0.878
0.7	0.786	0.698	0.813	0.899
0.8	0.791	0.696	0.816	0.910
0.9	0.779	0.708	0.814	0.901

Table 4: Relative frequencies, presence of select unigrams

Unigram	Relative frequency, anti-regime	Relative frequency, pro-regime
police	15.6	1.0
shelling	8.5	1.0
growing	7.7	1.0
leaders	6.7	1.0
activists	6.6	1.0
died	6.3	1.0
militants	5.8	1.0
dead	4.9	1.0
uprising	4.8	1.0
secretary	3.9	1.0
sarin	3.7	1.0
east	3.7	1.0

A ratio higher than 2.25 : 1.0 indicates the unigram is more frequent in anti-regime media. A ratio less than 2.25 : 1.0 indicates the unigram is more frequent in pro-regime media.

Table 5: Relative frequencies, absence of select unigrams

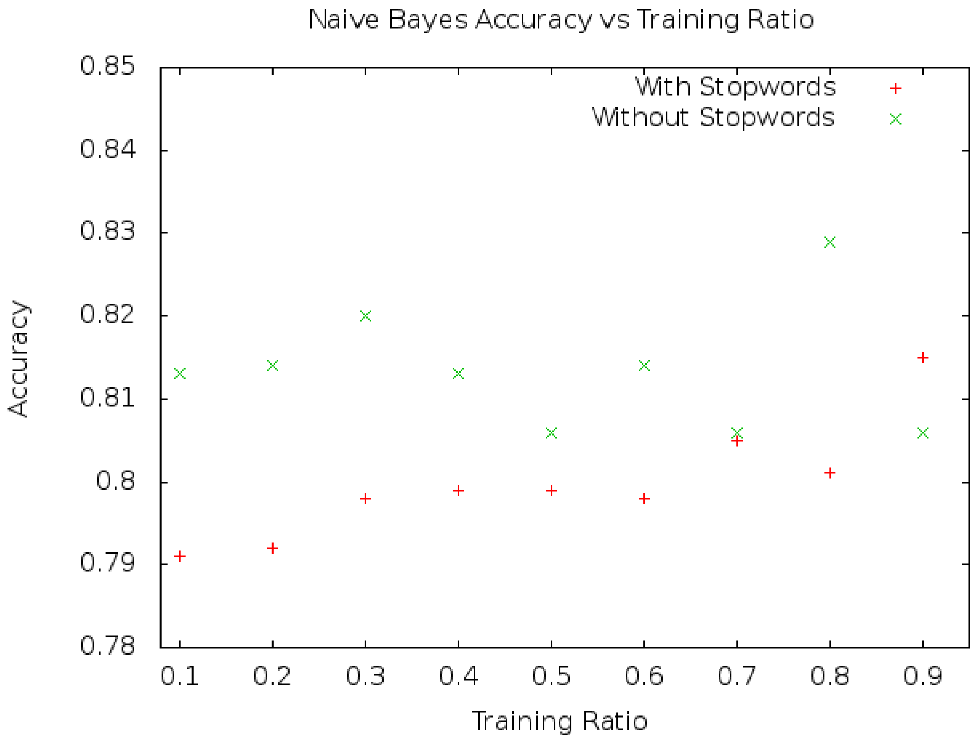
Unigram	Relative frequency, anti-regime	Relative frequency, pro-regime
not(people)	1.0	7.1
not(government)	1.0	2.8

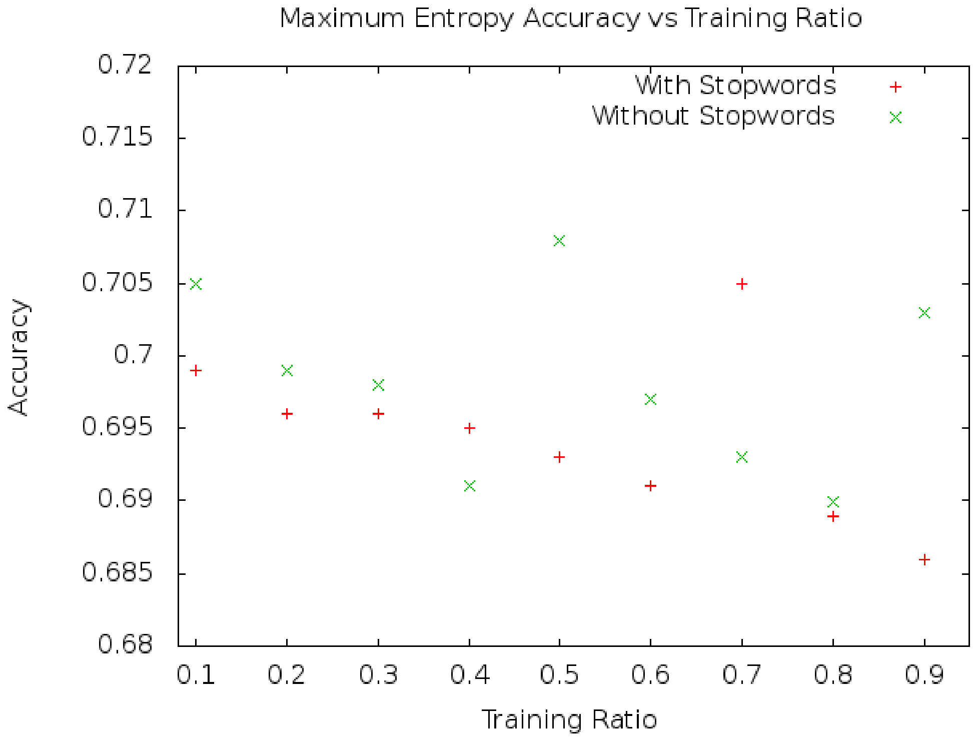
A ratio higher than 2.25 : 1.0 indicates the unigram is more frequently absent in anti-regime media. A ratio less than 2.25 : 1.0 indicates the unigram is more frequently absent in pro-regime media.

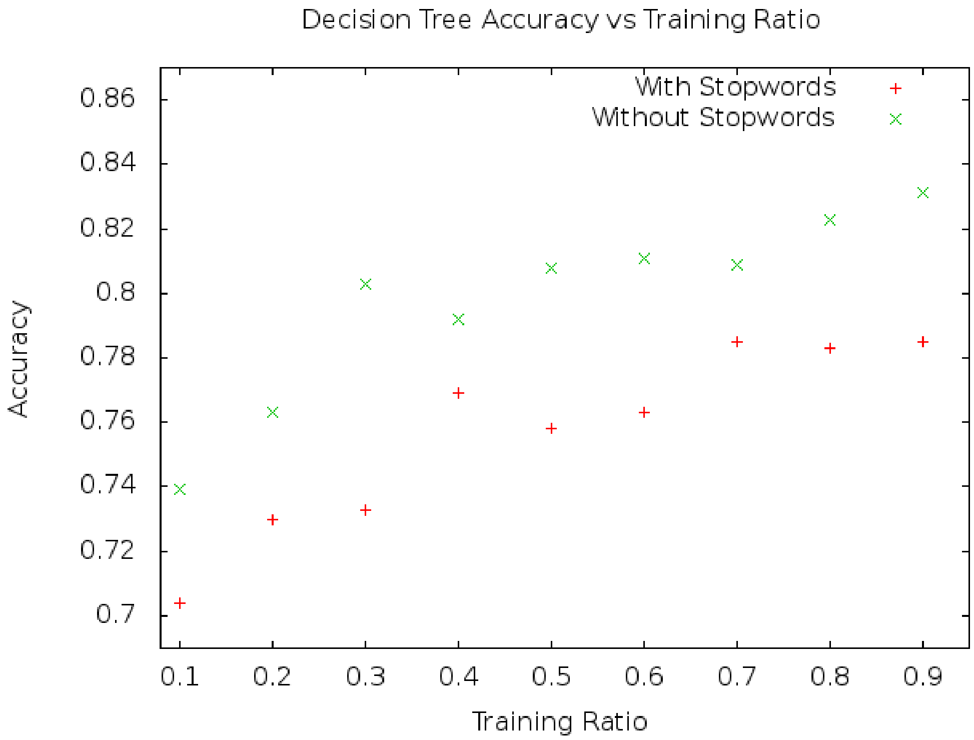
Table 6: Relative frequencies, presence of select bigrams

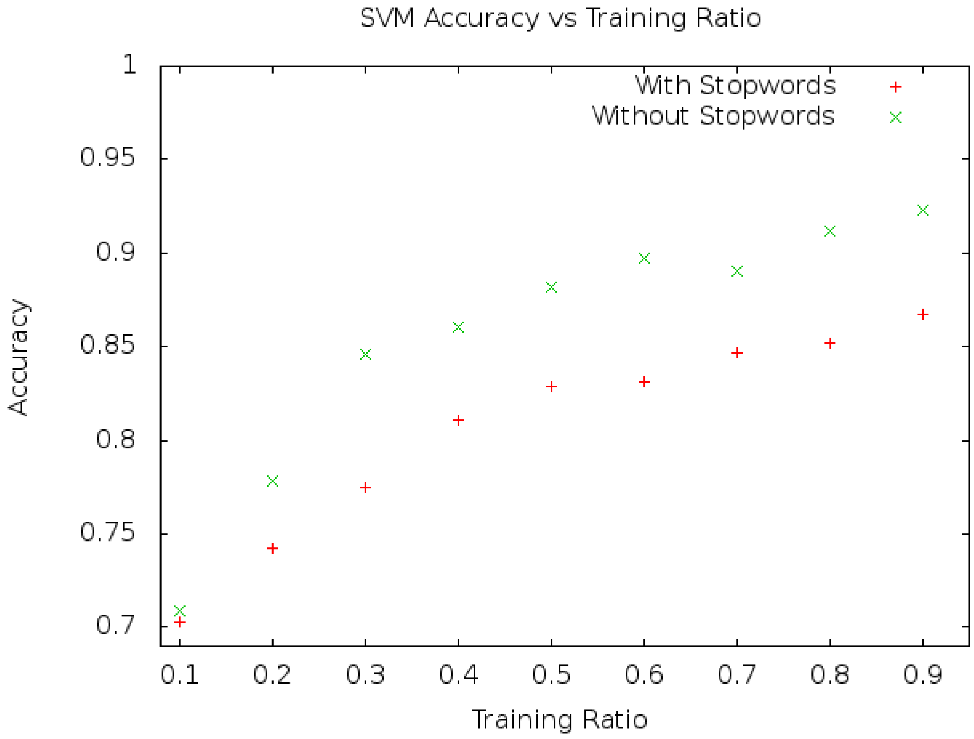
Unigram	Relative frequency, anti-regime	Relative frequency, pro-regime
to join	12.2	1.0
the site	11.8	1.0
the uprising	7.7	1.0
state department	6.9	1.0
secretary of	6.9	1.0

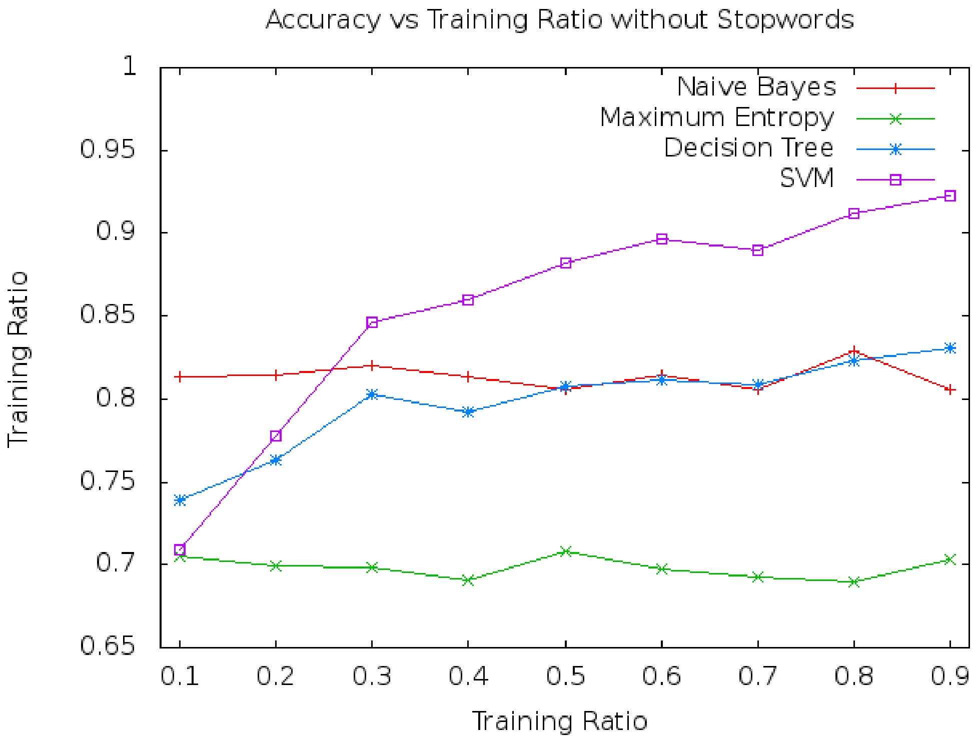
A ratio higher than 2.25 : 1.0 indicates the unigram is more frequent in anti-regime media. A ratio less than 2.25 : 1.0 indicates the unigram is more frequent in pro-regime media.

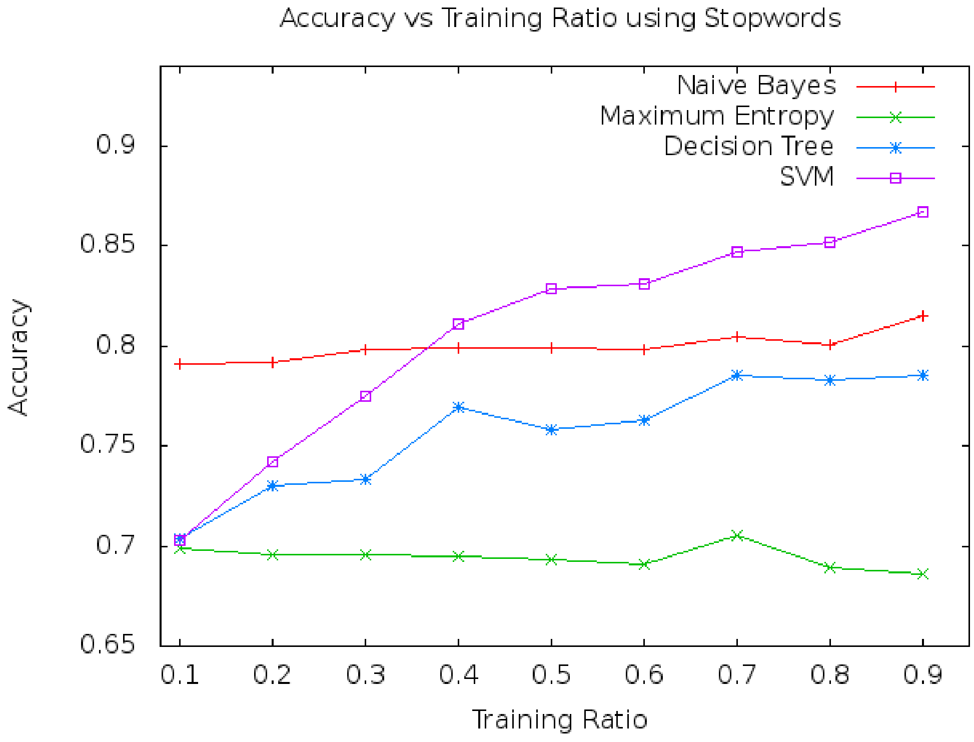

Naive Bayes Accuracy vs Training Ratio
Accuracy
Training Ratio
With Stopwords +
Without Stopwords ×
0.85
0.84
0.83
0.82
0.81
0.8
0.79
0.78
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

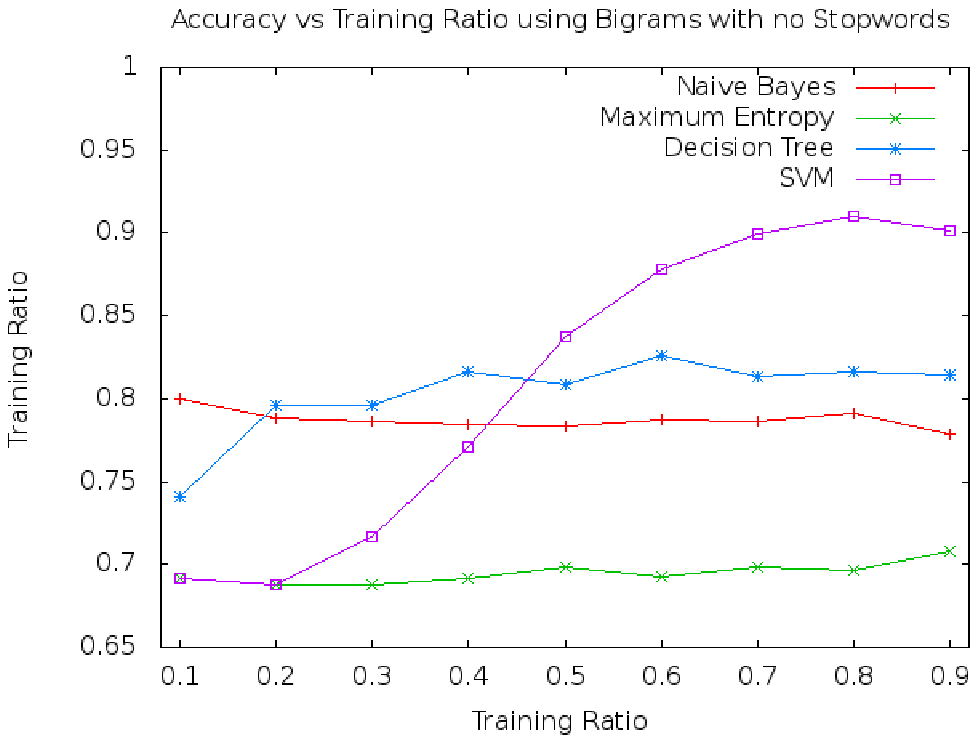

Maximum Entropy Accuracy vs Training Ratio
Accuracy
Training Ratio
With Stopwords +
Without Stopwords ×
0.72
0.715
0.71
0.705
0.7
0.695
0.69
0.685
0.68
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9


Decision Tree Accuracy vs Training Ratio
With Stopwords +
Without Stopwords ×
Accuracy
0.86
0.84
0.82
0.8
0.78
0.76
0.74
0.72
0.7
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Training Ratio


SVM Accuracy vs Training Ratio
With Stopwords +
Without Stopwords ×
Accuracy
1
0.95
0.9
0.85
0.8
0.75
0.7
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Training Ratio


Accuracy vs Training Ratio without Stopwords
Training Ratio
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Training Ratio
Naive Bayes
Maximum Entropy
Decision Tree
SVM


Accuracy vs Training Ratio using Stopwords
Accuracy
0.9
0.85
0.8
0.75
0.7
0.65
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
Training Ratio
Naive Bayes
Maximum Entropy
Decision Tree
SVM



Conclusion

Naive Bayes was the best method given limited training data, although SVM was a better method given more training data.

The vocabulary providing the best results in classification included both vocabulary related to the Syrian Civil War and vocabulary more commonly used by specific authors and news sources. It was found that focusing on vocabulary relevant to the Syrian Civil War was sufficient for outperforming ZeroR. The results show differences in lexicon used by pro-regime and anti-regime media, including words present in anti-regime media but absent in pro-regime media. Anti-regime media referenced unigrams such as ”police,” ”shelling,” ”growing,” ”leaders,” ”activists,” ”died,” ”militants,” ”dead,” ”uprising,” ”secretary,” ”sarin,” and ”east” far more frequently than pro-regime media. Pro-regime media omitted the unigram ”people” far more frequently than anti-regime media. There were also distinctions in bigram frequency.

Anti-regime media used the bigrams ”the uprising,” ”secretary of,” ”state department,” ”the site,” and ”to join” far more frequently than pro-regime media. The prevalence of ”secretary of” and ”state department” likely indicate that anti-regime media refered to the US response to events in Syria more frequently than pro-regime media.