Linguistic models for predicting stance in language

Subramanian Sivaraman ssivaraman@indiana.edu
Sharik Purkar sppurkar@iu.edu
Mithilesh Nanj minsrini@indiana.edu

***Abstract*—We present classification methods with diverse feature set for the task of classifying stance of tweets on pre-specified targets (atheism, Hillary Clinton, Feminist movement, climate change, Trump). Several prediction models were developed ranging from naive method to scikit random forest method varying by features used to classify. We explain the methodolgy followed in developing these models and compare the results of all the models. Memory based learning method Timbl was used for all the models with changes in feature sets (Subjectivity lexicon and arguing lexicon were also added along with Gold standard sentiment). Out of all the models, random forest model predicted the stance of Atheism in tweets with highest accuracy and F-measure.**
Keywords: Stance, Classification, Random Forest, Part of speech tagging, bag of words

I. Introduction

Stance detection is a problem of text classification. Unlike other text classification tasks, the goal is not to detect the sentiment, identify topics or authors of a text. Stance detection can be applied in the fields of behaviour analysis, political science, marketing analytics, web targeting analytics, etc. This problem deals with detecting stance of a user on a specified target. 5 targets were specified and a set of 2914 distinct tweets were taking in the training set. Three kinds of models were built with changes to feature sets. The first kind was developed by using POS tagging, bag of words and Timbl. Second kind used Random forest classifier in python scikit learn. Third kind was built using POS Tagging, MALT Parsing and Timbl. The test tweets was used to test and evaluate the model. In the random forest approach, 10-fold cross validation was used on the train set for training the model. All the three methods are shown in the flow diagrams fig 1, fig 2, fig 3 respectively. Every model developed is explained in separate sections in this paper. Results and comparison for all the models are shown in their respective sections.

II. Literature survey

JU NLP at SemEval-2016 Task 6: Detecting Stance in Tweets using Support Vector Machines gave us an idea about stance detection [1]. Other related works include DeepStance at SemEval-2016 Task 6: Detecting Stance in Tweets Using Character and Word-Level CNNs [2], ltl.uni-due at SemEval-2016 Task 6: Stance Detection in Social Media Using Stacked Classifiers [3], Any-Target Stance Detection on Twitter with Autoencoders [4] - all these describe in detail, methods of detecting stance in Twitter data.

III. Data

A. Data Source and Description

The data set given, contains tweets, their sentiments, six targets (Abortion, climate change, Hillary Clinton, feminist movement, Atheism, Trump) and stance class label (for, against, neutral) for training the model. Test data set contains tweets to be tested and the model is evaluated based on accuracy and f-score. Bag of words model, sentiment, MPQA subjectivity lexicon, Arguing lexicon, MALT parser output were used as features and the results are compared.

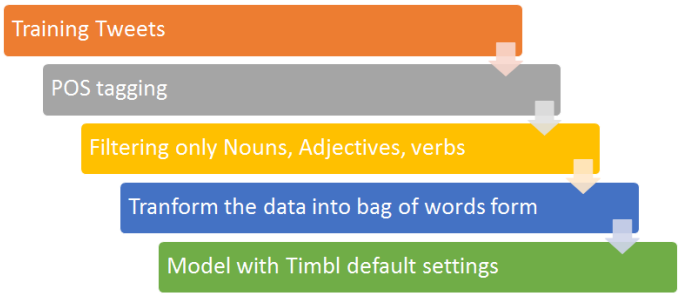

Fig. 1. Flow of Stance detection models - 1

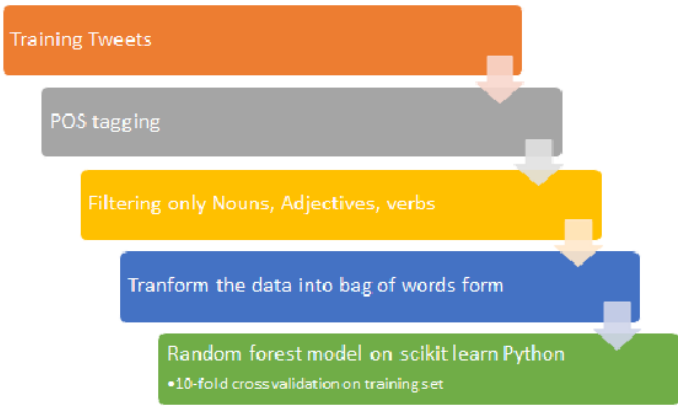

Fig. 2. Flow of Stance detection model - 2

B. POS Tagging and Bag of words

Part Of Speech tagging is the method of assigning parts of speech to each word in a text. Adjectives, nouns and verbs are considered as best representatives of tweets for this classification task since stance is usually highly correlated to words with one of these three parts of speech. Part of speech tagging is performed using NLTK in Python (tried two packages - treebank and perceptron, treebank tagged the words with greater accuracy). Stanford POS tagger is used to verify our tags. From the output, the verbs, adjectives and nouns are filtered. This can be done for all words together since POS tagging is independent of the targets.

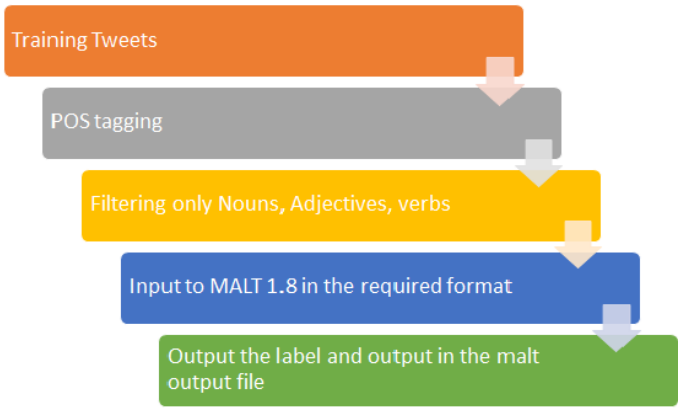

Fig. 3. Flow of Stance detection model - 3

TABLE I
TIMBL DEFAULT SETTINGS - RESULTS

Target	Accuracy	F-Measure
Atheism	39.5%	36.2%
Climate change	35.2%	24.3%
Hillary	38.4%	34.2%
Feminist movement	34.3%	28.9%
Trump	30.2%	35.2%

IV. Feature Extraction

All the words from POS tagging are taken as features and their values are 1 if the word is present in the tweet and 0 otherwise. This bag of words feature set along with the sentiment tag (observed high statistical correlation between sentiment and stance of a text) were taken as the features for prediction of stance. This basic set of features were then optimized and modified to improve the evaluation metrics.

V. Classification using Timbl

The constructed feature vector was used for classification using Timbl with default settings. The results are given in the table I for all the targets. The features and settings were then modified to improve the evaluation metrics of the model, using different methods which will be discussed in the next section.

VI. Feature optimization

Following steps were performed in order to optimize the features

- A basic threshold of 5 was set for all the features i,e. filtered out all features which had less than 5 occurrences in all the tweets
- Feature selection was performed based on Entropy of the features
- Features were ranked based on entropy and the top 70% of the features were retained
- This optimized or less sparse data set was used for training the model with modified settings on Timbl
- Hybrid version of decision tree and K-NN was used, since Cosine transform works well on sparse data set, Cosine was used as the distance metric, Information gain weighting was applied, k value was set to 3 The results with these settings for all targets are given in table II

TABLE II
FEATURE OPTIMIZATION - RESULTS

Target	Accuracy	F-measure
Atheism	41.7%	47.1%
Climate change	36.4%	38.5%
Hillary	39.1%	34.4%
Feminist movement	34.9%	30.9%
Trump	32.8%	36.7%

TABLE III
NAIVE METHOD - RESULTS

Target	Accuracy	F-measure
Atheism	22.7%	19.1%
Climate change	18.4%	21.5%
Hillary	19.2%	22.4%
Feminist movement	20.9%	26.9%
Trump	23.2%	17.7%

The results from this model show an increase in accuracy across all the targets. F-score of this model with respect to all targets also improve. This is primarily attributed to the hybrid model used instead of just a k-NN model.

VII. Naive method

Applying the naive method of including all words in the tweets as features without any pre-processing resulted in a much expected decline in the evaluation metrics. Including all features increased the sparsity and dimensions multifold and hence the model was trained with a lot of noisy data which led to a decline in accuracy and f-measure. The results are shown in table III

VIII. MPQA Subjectivity Lexicon

The MPQA subjectivity lexicon is a sample of gold standard words with type of subjectivity, polarity of words and the part of speech. Since it was only a sample of words, there were only a few occurrences of the lexicon in the tweets. Six features were added to the feature set - no. of positive words, no. of negative words, no. of strong subjective words and number of weak subjective words. Any of these features were included only if there was at least 5 occurrences. The results are shown in table IV

The results show a marginal increase in F-measures in all targets, but decrease in accuracies for Climate change and Feminist movement targets. It can be seen that adding these features improve F-measures.

TABLE IV
MODEL WITH MPQA LEXICON FEATURES - RESULTS

Target	Accuracy	F-measure
Atheism	42.9%	49.4%
Climate change	34.4%	47.2%
Hillary	40.3%	37.2%
Feminist movement	32.7%	34.9%
Trump	33.6%	36.2%

TABLE V
MODEL WITH ARGUING LEXICON FEATURES METHOD 1 - RESULTS

Target	Accuracy	F-measure
Atheism	38.1%	41.2%
Climate change	30.3%	38.9%
Hillary	36.0%	34.1%
Feminist movement	31.5%	28.1%
Trump	34.5%	29.9%

TABLE VI
MODEL WITH ARGUING LEXICON FEATURES METHOD 2 - RESULTS

Target	Accuracy	F-measure
Atheism	39.0%	43.1%
Climate change	32.3%	40.1%
Hillary	36.6%	36.9%
Feminist movement	31.9%	32.0%
Trump	33.9%	24.7%

TABLE VII
MODEL WITH ARGUING LEXICON FEATURES METHOD 3 - RESULTS

Target	Accuracy	F-measure
Atheism	44.7%	49.1%
Climate change	42.3%	45.5%
Hillary	41.1%	43.2%
Feminist movement	40.6%	39.4%
Trump	39.5%	36.8%

TABLE VIII
MODEL WITH MALT PARSER DEFAULT SETTINGS - RESULTS

Target	Accuracy	F-measure
Atheism	47.8%	51.4%
Climate change	40.9%	47.5%
Hillary	40.0%	36.4%
Feminist movement	32.7%	29.1%
Trump	38.7%	39.3%

IX. ARGUING LEXICON

The Arguing lexicon consists of categories of words and their sentiment in argument (gold standard). These words and their categories help us in predicting the direction of an argument about the tweet, for example intensifiers such as terribly, utterly express a negative sentiment about the target (there could be negators, but in general it is safe to assume). These words were included in the feature set by the following methods:

1) All the words present in the arguing lexicon were used as features as bag of words along with the already extracted features

   This increased the dimensionality and sparsity of the data set
2) Count of occurrences of all the words in each category

   In this method, categories were made the features and their values were count of occurrences. Also, it decreased the dimensionality of the data set when compared with the first data set
3) To further reduce dimensionality, category of the most frequently occurring words was taken as the value of the feature. The feature was named, 'Arguing category'

   If none of the words were occurring in the tweet, it will have a value of 'nil'. If there were equal occurrences from two or more different categories, then the category that appears more frequently across other tweets was taken as its value

   This was the most effective feature as it significantly improved the results when compared to other methods

Results from these three methods are shown in tables V , VI , VII

X. MALT PARSER

Malt parser is a data driven dependency parser which uses machine learning methods to parse data with the help of relationships in a graph database format. This improves the non-dependency parsing methods since the relationships from a graph database are used and also by use of machine learning methods such as SVM. Input with all words were sent in the format as required by the MALT parser version 1.8. Two models were trained, one using the default MALT parser settings, another model with customized settings.

1) With the default settings, the model was trained using talbanken05_train.conll -m learn

   This model yielded results comparable to that of the model with bag of words features + MPQA + Arguing lexicon
2) Settings were changed in MALT in order to improve the results - Cogniton parsing algorithm was used along with the large linear classifier instead of the default SVM classifier

   This model provided considerable improvements

Classification models were then trained and tested with outputs from these two parsers as features. Results from these models are shown in tables VIII , IX

XI. SCIKITLEARN - RANDOM FOREST

All these models were memory based. An ensemble based Random Forest classifier was built using scikit learn package in python 3.0 - 'sklearn.ensemble.RandomForestClassifier'. Since this builds n number of trees and averages all the results, it is expected to produce results better than memory learning

TABLE IX
MODEL WITH MALT PARSER CUSTOMIZED SETTINGS - RESULTS

Target	Accuracy	F-measure
Atheism	50.1%	54.4%
Climate change	40.3%	44.0%
Hillary	44.7%	45.1%
Feminist movement	39.6%	41.5%
Trump	36.5%	21.2%

TABLE X
RANDOM FOREST MODEL (25 TREES, GINI) - RESULTS

Target	Accuracy	F-measure
Atheism	61.4%	64.3%
Climate change	49.7%	52.6%
Hillary	47.1%	43.5%
Feminist movement	51.2%	37.9%
Trump	45.4%	44.9%

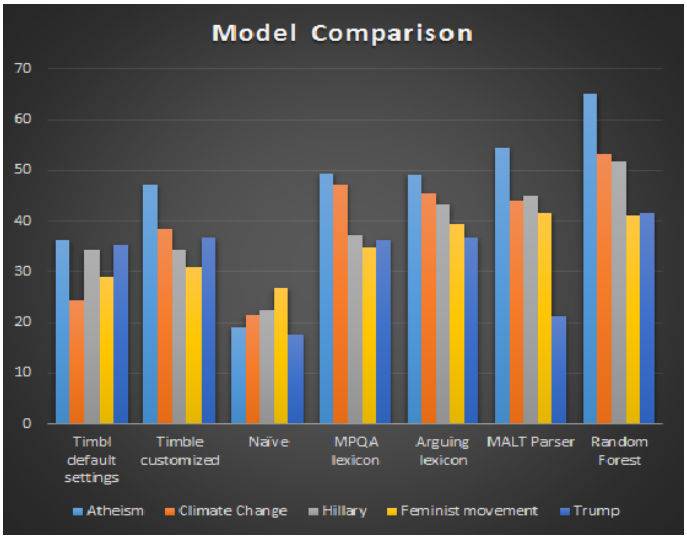

Fig. 4. Comparison of results

TABLE XI
RANDOM FOREST MODEL (100 TREES, ENTROPY) - RESULTS

Target	Accuracy	F-measure
Atheism	64.2%	65.1%
Climate change	51.4%	53.3%
Hillary	55.0%	51.9%
Feminist movement	50.9%	41.0%
Trump	46.5%	41.7%

methods. Two models were built varying the impurity index and the number of trees. Also, 10-fold cross validation was performed on the training data set to improve the training model.

1) The first model was built with n = 25 trees and the impurity was set to GINI
2) Second model was built with n=100 trees and the impurity was set to Entropy

Results from these models are given in tables X , XI

The results show that the F-measure is highest for the Random forest models.

XII. COMPARISON

This section compares the F-measures of all the models developed. Comparison chart is shown in the figure 4

XIII. ACKNOWLEDGEMENTS

This work was submitted as term paper in Computational Linguistics Class (Fall-2016) at Indiana University. We would like to extend our thanks and express our sincere gratitude to our adviser/Prof.Sandra Kubler for the continuous support and guidance through out the period of this course.

XIV. CONCLUSION

This project primarily deals with building learning models for Stance prediction in language with targets that are specified beforehand. Stance prediction can be used and applied in social networking, customer behaviour analysis, socio economic and political science and many other fields. This prompts us to build efficient and high precision models and find creative methods for feature extraction. In our project, we got a good exposure to most of the methods that are being used and we also tried to improvise on that by tuning parameters and deriving features.

Nancy Green, Kevin Ashley, Diane Litman, Chris Reed, and Vern Walker, editors. 2014. Proceedings of the First Workshop on Argumentation Mining. Association for Computational Linguistics, Baltimore, Maryland, June

REFERENCES

[1] Pranav Anand, Marilyn Walker, Rob Abbott, Jean E. F.Tree, Robeson Bowmani, and Michael Minor *Cats rule and dogs drool!: Classifying stance in online debate*,In Proceedings of the 2nd workshop on computational approaches to subjectivity and sentiment analysis, pages 19

[2] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. *Rich feature hierarchies for accurate object detection and semantic segmentation*, In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 580587

[3] Nancy Green, Kevin Ashley, Diane Litman, Chris Reed, and Vern Walker, editors. 2014 *Proceedings of the First Workshop on Argumentation Mining. Association for Computational Linguistics, Baltimore, Maryland, June*

[4] Luciano Barbosa and Junlan Feng, *Robust sentiment detection on twitter from biased and noisy data*. Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pages 3644. Adam Bermingham and Alan Smeaton. 2010.