# IUCL at SemEval-2016 Task 6: An Ensemble Model for Stance Detection in Twitter

**Can Liu, Wen Li, Bradford Demarest, Yue Chen, Sara Couture, Daniel Dakota, Nikita Haduong, Noah Kaufman, Andrew Lamont, Manan Pancholi, Kenneth Steimel, Sandra Kübler**
Indiana University
{liucan,wl9,bdemares,yc59,scouture,ddakota,nhaduong,nokaufma,
alamont,mpanchol,ksteimel,skuebler}@indiana.edu

## Abstract

We present the IUCL system, based on supervised learning, for the shared task on stance detection. Our official submission, the random forest model, reaches a score of 63.60, and is ranked 6th out of 19 teams. We also use gradient boosting decision trees and SVM and merge all classifiers into an ensemble method. Our analysis shows that random forest is good at retrieving minority classes and gradient boosting majority classes. The strengths of different classifiers wrt. precision and recall complement each other in the ensemble.

## 1 Introduction

Stance detection is a difficult task since it often requires reasoning in order to determine whether an utterance is in favor of or against a specific issue. In the shared task (see Mohammad et al. (2016) for details about the shared task), we interpret it as a variant of sentiment analysis and adopt an approach that combines shallow lexical features with an ensemble of different supervised machine learning classifiers. Previous work has shown that using "arguing" features based on an arguing lexicon along with modal verbs and targets identified via syntactic rules (Somasundaran and Wiebe, 2010); finding polarized relations between aspects and topics (Somasundaran and Wiebe, 2009); adding semantic frames (Hasan and Ng, 2013) and contextual features (Anand et al., 2011) generally improve results. Since some of these features do not generalize across targets (Anand et al., 2011), and since we have an additional challenge in processing Twitter data, we rely on unigram features and word vectors. This means that our approach is incapable of handling sarcasm or humor. Instead, it provides a robust basis on which we can later add more informative features.

Our approach consists of classifiers with a bag of words (unigrams) or with word vectors as features. We use three separate classifiers (SVMs, random forest, gradient boosting decision trees) and an ensemble classifier (TiMBL). Our official submission is the random forest classifier with word unigrams.

## 2 Methods

We use the data sets provided by the SemEval-2016 shared task 6 (Mohammad et al., 2016).

### 2.1 Preprocessing

Preprocessing mostly consists of tokenization. During tokenization, we normalize capitalization, and all punctuation signs are separated except for @ and #, as these symbols indicate hashtags and handles. We extract frequency counts of each token in the entire corpus and in each stance (Favor, Against, None) per target for use in the feature selection process.

We experimented with TWEEBOPARSER (Kong et al., 2014), a dependency parser specifically designed for Twitter data, to extract dependency relations among words. We extract POS tags, multiword expressions, and dependency triples from the parses. However, due to the feature sparsity, none of them improved over unigrams. Thus, they are not used in the final systems.

### 2.2 Features

One of the major decisions in developing a machine learning system for stance detection lies in



*Proceedings of SemEval-2016*, pages 406–412,
San Diego, California, June 16-17, 2016. 

| Model | Features |
|---|---|
| GBDT | GloVe word vectors |
| random forest | unigrams + IG |
| SVM | unigrams + IG |
| ensembleG | three classifiers + global |
| ensembleNG | three classifiers only |

**Table 1:** Summary of features for each model. The random forest model constitutes our official submission.

the choice of features and of feature representations. Detecting stance in political tweets can be regarded as a form of sentiment analysis for short text, and we assume that different stances of tweets are partially expressed by the choice of words. For example, not mentioning any words that express a polarized attitude indicates that a tweet is most likely a None stance. Tweets are relatively short documents, we use bag of words (unigrams) since in this case bigrams and trigrams are likely to be too sparse to be informative. Another possibility would be to follow approaches in sentiment analysis and use sentiment lexicons. However, such lexicons are normally general purpose resources, and domain specific information is not included. In contrast, we need such domain specific knowledge, for example to capture the fact that "dear lord" is an indication of a negative stance towards the target Atheism while it may have a different meaning when it occurs for the target Hillary Clinton. Since unigrams include a high number of irrelevant features and also constitute a rather impoverished representation, we use feature selection as well as word vectors in our experiments.

Table 1 summarizes the features used for each of our models. We use information gain (IG) for feature selection on unigrams. Global refers to global features (see section 2.2.3). The three classifiers are GBDT, random forest, and SVM; the ensemble uses their output (predicted label and its probability).

#### 2.2.1 Feature Selection

There are issues resulting from the large number of bag-of-words features: 1) Not all words are good indicators for stance; some words occur evenly across the data set. 2) Rare words, which are less likely to occur in the test data, do not contribute much. To alleviate these problems, we perform feature selection using information gain (IG). IG estimates the amount of information a word gives for the decision on the stance. We choose IG because it has been shown to be robust across different sentiment analysis data sets and across different skewing ratios, compared to other feature selection methods (Liu et al., 2014). Note that different from its use in decision trees, we use IG as an external filter to select a subset of features, before and independent of any classifiers.

#### 2.2.2 Word Vector Features

One limitation of bag-of-words features is that they are very sparse, and they cannot handle out-of-vocabulary words properly. Since tweets are relatively short, and the amount of official training data is small, it is likely that the out-of-vocabulary rate is high. Thus we also build models using word vectors, which represent each word with a vector of continuous values. Word vectors have been shown to capture the similarity among words and thus alleviate data sparseness (Collobert et al., 2011).

We have experimented with two different word vector models, word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014). We have used the pre-trained word2vec obtained from the Google News dataset, which contains a 300-dimensional vector representation for 3 million words and phrases[1], and the pre-trained GloVe, which is obtained from 2 billion tweets and has a 250-dimensional vector representation for 1.2 million words and phrases[2].

To construct a representation for a tweet, we look up a word in the word vectors model, then average all vectors for words to produce a vector representation for the tweet. For example, to represent a 15 word tweet using word2vec, we first obtain a 300-dimensional vector for each word, then average all 15 vectors. This means that the word order is lost and the representation constitutes a "bag of vectors".

**Comparing Word Vectors** We have performed a comparison of both word vector variants in a 5-fold cross validation experiment on the training data. Table 2 summarizes the results. We can see that GloVe performs consistently better than word2vec except for Feminist where word2vec is 0.6% better than

[1] https://code.google.com/archive/p/word2vec/

[2] http://nlp.stanford.edu/projects/glove/

| Target | Word2vec | GloVe |
|---|---|---|
| Abortion | 61.4 | 62.4 |
| Atheism | 62.6 | 66.4 |
| Climate | 69.9 | 71.1 |
| Feminist | 53.8 | 53.2 |
| Hillary | 59.5 | 61.1 |

**Table 2:** Comparing word2vec and GloVe.

GloVe. We assume that this performance gap is mainly caused by the domain difference from which the word vectors are obtained: We used GloVe pre-trained on tweets and word2vec pre-trained on news. This leads to a higher number of out-of-vocabulary words for the word2vec model. In other words, GloVe provides a broader coverage for this data set.

#### 2.2.3 Global Features

The bag-of-words features used in the classifiers (see section 3) assume that the words are considered independently. However, in many situations, it is the distributions of positively and negatively oriented words that determine the final stance of a tweet. A low coverage of words from these two distributions is a strong indicator for None stance as well. This is especially important for the ensemble classifier. For this reason, we have developed two additional features for the ensemble, which capture information from these two distributions: one feature for positive orientation and one for negative orientation. The feature is a numeric score, representing the association of a tweet with positive or negative stance respectively. The positive orientation is calculated based on the following equation:

$$score_T^{pos} = \frac{1}{|T|} \sum_{w \subset T} \frac{freq(w) \text{ in POS}}{\sum_{w' \subset V} freq(w') \text{ in POS}}$$

where $T$ is a tweet, $|T|$ is the tweet length excluding stop words. $V$ is the entire vocabulary. $Freq(w)$ is the frequency count of $w$ in the following set. $POS$ is the set of all positive tweets. This score measures for each word (its lemma) the association with positive stance, sums up all words in the tweet, and normalizes the score by the tweet length. The score for the negative orientation is calculated accordingly.

The None orientation is not calculated since it is already represented by the absence of positively or negatively oriented words. I.e., we assume that if a tweet has low positive and negative orientations, it indicates a None stance.

### 2.3 Adding Manually Annotated Data

We mined additional tweets for each of the five targets in Nov. 2015 by searching for hashtags relevant to the targets. These tweets are not included in the final systems since they increased the class imbalance. We will investigate better options for including the data in the future. Hashtags for Abortion include #abortion, #abortionrights, and #prolife; Atheism includes #atheism, #atheist, and #theist; Climate includes #actionclimate and #climatechange; Feminist includes #feminism, #feminist, #heforshe, and #womensrights; and Hillary includes #HillaryClinton.

Tweets were then annotated for stance, following the guidelines used for the annotation of the official shared task data[3]. Two annotators participated in the annotation process. The number of additional tweets ranged between 260 and 2,400 per target.

## 3 Classifiers

Since there is little research on determining the best fitting bias for stance detection, we explore three different classifiers for the stance classification, support vector machines (SVM), random forest, and gradient boosting decision trees (GBDT). For all three classifiers, we use the implementations in Scikit-Learn (Pedregosa et al., 2011).

We choose SVM because it is the most widely used machine learning model for text classification and sentiment analysis (e.g., (Pilászy, 2005)).

Additionally, it has been shown to be robust with high dimensional features (e.g., (Joachims, 1998)). Random forest is adopted because of its capability of reducing overfitting by performing sampling on data points and on feature subspaces. GBDT is selected because it works well with continuous numerical features such as word vectors.

We train individual classifiers for each target. Parameters are optimized in a 5-fold cross-validation over the training data. SVM and random forest are trained on different numbers of selected unigrams

[3]See http://alt.qcri.org/semeval2016/task6/data/uploads/stance-question.pdf.

for each target: 1,700 for Abortion, 1,535 for Atheism, 1,381 for Climate, 1,749 for Feminist, and 1,704 for Hillary. GBDT is trained on the word vectors: 300 dimensions for word2vec and 250 dimensions for GloVe. Additional experiments are performed with a standard feed-forward neural network on word vectors. These showed better performance on the training set for some targets, but overall, GBDT prove to be more reliable.

**SVM** Our initial experiments using cross validation on training data showed that linear kernel performed better than non-linear ones, and that the LinearSVC implementation (one-vs-rest strategy for multi-class) outperformed SVC (one-vs-one strategy). The optimal parameters differ for each target: 0.015-0.3 for the slack variable; standard hinge or squared hinge for the loss function; and L2 norm for the penalty term.

**Random Forest** The parameters for random forest are: 50, 70, or 90 for the number of trees; 500 or All for the number of features to consider when looking for the best split; 200, 500, or unlimited for the maximum depth of trees.

**GBDT** The gradient boosting decision trees (GBDT) classifier is used in combination with word vector features. Our initial experiments showed that GBDT handles word vector features better than SVM and random forest. The optimal parameter range for different targets are: 80-100 for number of estimators; 0.05-0.3 for learning rate; false for warm start; and 0.5-1.0 for subsample ratio.

**Ensemble Classifier** Since initial experiments with the three classifiers showed considerable differences across targets and stances, we investigate whether an ensemble classifier would benefit from aggregating their predictions. For the ensemble classifier, we choose a memory-based learner, TiMBL, because of the need to operate on a small set of rather abstract features: stance predictions and confidence scores from the three classifiers along with the global features (see section 2.2.3).

We use TiMBL (Daelemans et al., 2009) version 6.4.2, and perform 5-fold jackknifing to generate the training set for this ensemble classifier. Parameter optimization is performed on the five folds. The best parameters are different in each target: 7-29 for the number of neighbors; default minority voting for class voting in most cases; Modified Value Distance, Jeffrey divergence, and cosine distance for distance metric; and gain ratio for feature weight in most cases.

| Team | Official Metric |
|---|---|
| MITRE | 67.82 |
| IUCL-RF | 63.60 |

**Table 3:** Official results of the IUCL-RF system in comparison to the best system.

| Model | Official Metric |
|---|---|
| GBDT | *64.64* |
| Random Forest | 63.60 |
| SVM | 61.93 |
| EnsembleG | 62.46 |
| EnsembleNG | **66.14** |

**Table 4:** Overall comparison of all IUCL systems. The best accuracy of an individual classifier is shown in italics, the best overall result in bold.

## 4 Results

### 4.1 Official Result

Since the ensemble classifier was not completed in time for submission, we had to decide which individual classifier to submit. The random forest model is selected based on a five-fold cross validation on the training set. This system reaches a score of 63.60 (macro-averaged F), as shown in table 3, the sixth best result out of 19 participating systems. This result is approximately 4 percent points lower than that of the highest performing system.

### 4.2 Additional Results

#### 4.2.1 Overview of All Classifiers

Table 4 shows the results of the three individual classifiers as well as of the two ensemble model variants, one combining only the individual classifiers' outputs (EnsembleNG), the other one (EnsembleG) including also the global features (see section 2.2.3). These results show that the GBDT approach using GloVe reaches the highest result (64.64) among the individual classifiers. The random forest classifier, which constitutes our official submission is about 1 percentage point lower (63.60), and the

| | Abortion | | | | | | | | Atheism | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Favor | | Against | | None | | | | Favor | | Against | | None | |
| Model | Acc | F | Prec | Rec | Prec | Rec | Prec | Rec | Acc | F | Prec | Rec | Prec | Rec | Prec | Rec |
| GBDT | *65.0* | 53.6 | 52.6 | 21.7 | 75.6 | 77.2 | 38.2 | 57.8 | 67.3 | 56.4 | 37.0 | 31.2 | 82.3 | 75.6 | 37.0 | 60.7 |
| RF | *65.0* | 57.6 | 43.6 | 37.0 | 83.8 | 68.3 | 41.4 | 80.0 | ***70.5*** | 57.9 | 45.0 | 28.1 | 81.2 | 81.2 | 40.0 | 57.1 |
| SVM | 60.7 | 58.6 | 43.6 | 52.2 | 81.6 | 60.8 | 36.9 | 68.9 | 59.1 | 51.9 | 26.1 | 37.5 | 81.5 | 66.2 | 27.0 | 42.9 |
| EnsembleG | 62.9 | 46.3 | 55.6 | 10.9 | 75.1 | 73.5 | 37.2 | 71.1 | 69.1 | 45.9 | 50.0 | 6.2 | 76.1 | 85.6 | 36.1 | 46.4 |
| EnsembleNG | **66.8** | 60.2 | 50.0 | 39.1 | 80.2 | 73.0 | 43.1 | 68.9 | 69.1 | 50.6 | 57.1 | 12.5 | 74.0 | 88.7 | 28.6 | 21.4 |
| | Climate | | | | | | | | Feminist | | | | | | | |
| | | | Favor | | Against | | None | | | | Favor | | Against | | None | |
| Model | Acc | F | Prec | Rec | Prec | Rec | Prec | Rec | Acc | F | Prec | Rec | Prec | Rec | Prec | Rec |
| GBDT | ***72.8*** | 41.8 | 82.0 | 85.4 | 0.00 | 0.00 | 43.9 | 51.4 | *57.9* | 51.6 | 30.3 | 34.5 | 69.3 | 72.7 | 44.4 | 27.3 |
| RF | 68.0 | 39.1 | 82.7 | 74.0 | 0.00 | 0.00 | 40.7 | 68.6 | 57.2 | 51.1 | 31.1 | 39.7 | 73.9 | 61.7 | 46.6 | 61.4 |
| SVM | 68.6 | 39.8 | 79.7 | 79.7 | 0.00 | 0.00 | 39.1 | 51.4 | 55.4 | 54.6 | 33.9 | 67.2 | 76.5 | 55.2 | 47.4 | 40.9 |
| EnsembleG | 69.2 | 39.6 | 81.2 | 77.2 | 0.00 | 0.00 | 42.3 | 62.9 | **65.6** | 44.9 | 57.1 | 6.9 | 68.9 | 88.5 | 48.8 | 47.7 |
| EnsembleNG | 72.2 | 40.5 | 84.2 | 78.0 | 0.00 | 0.00 | 47.3 | 74.3 | 62.8 | 57.9 | 39.4 | 44.8 | 75.1 | 72.7 | 47.6 | 45.5 |
| | Hillary | | | | | | | | | | | | | | | |
| | | | Favor | | Against | | None | | | | | | | | | |
| Model | Acc | F | Prec | Rec | Prec | Rec | Prec | Rec | | | | | | | | |
| GBDT | 64.4 | 48.7 | 40.0 | 13.3 | 66.0 | 93.6 | 63.9 | 29.5 | | | | | | | | |
| RF | ***70.2*** | 49.8 | 75.0 | 13.3 | 70.5 | 84.9 | 68.8 | 70.5 | | | | | | | | |
| SVM | 62.0 | 55.3 | 36.8 | 46.7 | 70.2 | 68.6 | 62.9 | 56.4 | | | | | | | | |
| EnsembleG | 63.4 | 44.1 | 100.0 | 8.9 | 66.0 | 79.1 | 55.3 | 60.3 | | | | | | | | |
| EnsembleNG | 67.8 | 51.6 | 80.0 | 17.8 | 71.1 | 77.3 | 60.2 | 75.6 | | | | | | | | |

**Table 5:** Detailed comparison. Best accuracies of individual classifiers are shown in italics, best overall results in bold. (F = macro-averaged F over Favor and Against; official score.)

SVM classifier is about 1.5 percentage points below that (61.93). A closer look at the ensemble variants shows that using the global features has a detrimental effect across all targets, most likely because this information is too coarse. The other ensemble classifier improves over GBDT by 1.5 percentage points (66.14). This shows that we can benefit from important information from all individual classifiers.

### 4.2.2 Further Analysis

While the official scorer averages the results over all five targets, we are interested in whether our classifiers show a stable performance across targets, and why the ensemble model benefits from combining all individual classifiers. For this reason, we modified the scorer so that it would calculate accuracy, precision, and recall for individual stances per target separately. The results are shown in table 5. The official metric is the macro-averaged F-measure on Favor and Against while accuracy is equivalent to the micro-averaged F-measure based on all classes.

The results show a more diverse picture: For the individual classifiers, GBDT reaches the highest accuracies for the targets Climate and Feminist, random forest for Atheism and Hillary, and they tie for Abortion. For the ensembles, the version without global features reaches higher accuracies for Abortion, Climate, and Hillary, the version with global features has a higher accuracy for Feminist, and they tie for Atheism.

EnsembleNG, which reaches the best score across all targets, only reaches the best score for two targets: Abortion and Feminist. It reaches lower results than the best individual classifier for 3 targets: Atheism, Climate, and Hillary. However, since the best results for the latter 3 targets are reached by different individual classifiers (random forest for Atheism and Hillary; GBDT for Climate), we assume that the ensemble provides the best compromise.

In order to obtain a better understanding of the differences in performance of classifiers across targets, we have analyzed the distribution of stances per target. Table 6 shows the distribution in training and test data. If we combine the information from table 5 with the stance distributions, we notice that a major advantage of the random forest classifier is its

| Data Set | Stance | Abortion | Atheism | Climate | Feminist | Hillary |
|---|---|---|---|---|---|---|
| Train | Favor | 18 | 18 | 54 | 32 | 17 |
| | Against | 55 | 59 | 4 | 49 | 57 |
| | None | 27 | 23 | 42 | 19 | 26 |
| Test | Favor | 17 | 14 | 73 | 20 | 15 |
| | Against | 67 | 73 | 7 | 64 | 58 |
| | None | 16 | 13 | 20 | 16 | 27 |

**Table 6:** Class distribution across targets in percentage.

high recall on the None stance, which is generally (one of) the minority class(es). For the second minority class (Favor for Abortion, Atheism, Hillary, and Feminist; and Against for Climate), the picture is less clear: For Climate, none of the classifiers manage to identify any of the Against tweets. For Abortion and Feminist, random forest also shows a high recall for Favor, but for Atheism and Hillary, its precision is considerably higher. In contrast, GBDT reaches a higher recall for the majority class (with Atheism as the only exception). SVM generally has precision and recall values between or below the other classifiers. The only exception is the target Feminist, where SVM reaches the highest precision for all three stances.

One hypothesis that could be drawn from the analysis above is that the GBDT model is better suited for finding examples of the majority classes while random forest is better at finding minority class examples. However, when we compare the targets Abortion and Atheism, the class distribution is similar, but the performance of the two classifiers is vastly different: For Abortion, GBDT reaches higher recall for the majority class (Against) and higher precision for Favor. For Atheism, it has a higher precision for the majority class and a higher recall for Favor. The reasons for these different behaviors need to be determined in future work.

## 5 Conclusion

In this shared task, we regard stance detection as a special case of sentiment analysis, using supervised classifiers and bag of unigrams and word vectors as features. Our submitted system is based on a random forest classifier because of its capability to handle overfitting and to generalize over the test data. Since the amount of available training data is small, random forest's ability to sample data points and feature subspaces reduces data sparsity. The submitted system has an official score of 63.60 and ranked 6th out of 19 teams.

We also experimented with other single models (SVM and GBDT) and with an ensemble model built on a memory-based classifier. The GBDT model using GloVe word vectors reaches a higher score of 64.64, which may be a result of the word vectors' capability to capture similarities among words, which helps in dealing with out-of-vocabulary words. The ensemble model that aggregates information from the three individual classifiers reaches the highest performance of 66.14. Our hypothesis is that different strengths (e.g., good performance for minority/majority classes) from individual models complement each other in the ensemble.

However a closer look at the performance of all classifiers and ensembles across individual targets shows that no system reaches consistently good results across all targets. The best performing ensemble (EnsembleNG) outperforms individual classifiers only for Abortion and Feminist; for the other targets, random forest or GBDT reach higher accuracies. Some of the variation in system performance can be explained by the class imbalance present in the data sets for the different targets, but further work is required to identify other factors.

Finally, it is worth pointing out that our approach to stance detection utilizes very surface oriented features. To boost performance, we may need to develop methods that incorporate inference, entailment, and world knowledge, for example, to handle cases such as "keep H. out of the white house".

## Acknowledgement

This work is based on research supported by the U.S. Office of Naval Research (ONR) via grant #N00014-10-1-0140.

## References


Pranav Anand, Marilyn Walker, Rob Abbott, Jean E Fox Tree, Robeson Bowmani, and Michael Minor. 2011. 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 1–9, Portland, OR.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2009. TiMBL: Tilburg memory based learner – version 6.2 – reference guide. Technical Report ILK 09-01, Induction of Linguistic Knowledge, Computational Linguistics, Tilburg University.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing (IJCLNP 2013)*, pages 1348–1356, Nagoya, Japan.

Thorsten Joachims. 1998. *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. Springer.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A Smith. 2014. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 1001–1012, Doha, Qatar.

Can Liu, Sandra Kübler, and Ning Yu. 2014. Feature selection for highly skewed sentiment analysis tasks. In *Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP)*, pages 2–11, Dublin, Ireland.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval'16, San Diego, CA, June.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar.

István Pilászy. 2005. Text categorization and support vector machines. In *Proceedings of the 6th International Symposium of Hungarian Researchers on Computational Intelligence*, Budapest, Hungary.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing stances in online debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 226–234, Suntec, Singapore.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124, Los Angeles, CA.