Classifying Stance Using Profile Texts

Anonymous ACL submission

Abstract

This paper discusses our attempts to classify the stance of tweets in the context of an ongoing effort to assess the Social License to Operate (SLO) of mining companies, where SLO is a measure of the company's level of support from their constituencies. Our prototype system deploys an Support Vector Machine (SVM) classifier that: (1) relies on the rule-based coding of tweets for the training set, which allows for the construction of larger training and testing sets without resorting to the error-prone and expensive practices of manual coding or crowd-sourcing; and (2) includes Twitter author profile texts as input features, which provides the classifier with more data on the author's public views. Both practices are shown to be effective and are recommended for use in stance classification in other SLO domains.

1 Introduction

Following the pattern of the SemEval-2016 Task 6 challenge (Mohammad et al., 2016), in which systems attempted to classify the stance of tweets with respect to targets such as controversial political issues and opposing political candidates, our work attempts to use tweets to determine the Social License to Operate (SLO) of mining companies doing work in [country][1] (Gunningham et al., 2004). SLO is a measure of the public's support for a company or its projects. This is important information to distill, both for the target company's marketing and strategic decision making, as lack of support (or low SLO), can result in projects being cancelled, and for government organizations setting public policy.

This paper discusses our ongoing effort to build a classifier that monitors the stance of mining-related tweets collected by real-time Twitter feeds. We review relevant work, describe our training and testing datasets of Twitter data, detail the structure of our baseline Support Vector Machine (SVM) classifier, and review the results of our training and testing. The key challenges we have faced in this work are: first, to collect sufficient training data, particularly for pro-mining-company tweets, which are relatively rare; and second, to find sufficient input data in the tweet stream, beyond the raw tweet texts themselves, which didn't prove sufficient to train an accurate classifier. Our conclusions are that it has been possible to identify rules that can be used to reliably auto-code tweets as stance for, against or neutral for a sufficiently large training set, and that the inclusion of author profile texts as input features to a baseline SVM classifier, in addition to the raw tweets themselves, significantly improves the accuracy of the classification. We believe that these techniques would be useful for classifying SLO in other domains, beyond mining.

2 Related Work

Using the SemEval-2016 Task 6 challenge (Mohammad et al., 2016) as a model, our work has focused on automatically coding tweets as stance for, against and neutral with respect to target mining companies (e.g., Adani, BHP, RioTinto). In this mining domain, however, stance positions do not necessarily have the zero-sum context inherent in SemEval-2016's political domain. That is, support for one mining company doesn't necessarily imply rejection of other mining companies.

To allow more direct comparison with the SemEval-2016 results, our work adopts the same baseline SVM classifier and the same macro-averaged F1 metric. However, our approach differs in two key ways.

First, we deployed rule-based coding to build

[1] Removed to preserve anonymity.

the training set. SemEval-2016 Task 6 identified a set of what they call query hashtags, some of which tend to collect tweets that are stance for, others to collect tweets that are stance against and yet others to collect tweets that are stance neutral. They used these query hashtags to build a dataset and then manually-coded those samples for use in training and testing. Due to the time required to do manual coding, this led to somewhat smaller training sets (i.e., 400–700 tweets for each of their five task A targets), which, in turn, led to somewhat less effective training. This helps to explain why the SemEval baseline SVM classifier tended to out-perform the submitted deep neural networks, which tend to require more data. In our work, we collected query hashtags as well, using them to identify tweets about the mining companies, but we also collected a sub-set of those hashtags that reliably identified stance for, against and neutral tweets. We then used this sub-set, which Mohammad et al call stance-indicative hashtags, to create auto-coding rules to build our training set. The resulting training set is somewhat larger (i.e., 1863 tweets on Adani), split evenly between stance-for, against and neutral tweets, and labelled with sufficient accuracy. This approach allowed us to avoid the difficulty of achieving strong intercoder agreement on crowd-sourced coding tasks.

Second, we added the Twitter user profile text as an input feature to the classifier. The SemEval-2016 work did not include these profile texts in its datasets and, we believe as a consequence, included some miscodings in its manually-coded, gold standard testing set.[2] While it was sensible for SemEval-2016 to ignore these texts, which allowed it to establish a baseline for future work on stance classification, we believe that including text in which the author describes themself and their beliefs is an important improvement for stance detection systems using Twitter data. We are not aware of other work that has used these profile texts as input features.

3 Datasets

Raw Dataset: The dataset includes 684,640 raw tweets with associated author profile description texts, collected daily from the Twitter API from January, 2010 through May, 2018. It was restricted to tweets: (1) posted from [Country], based on the tweet timezone, which was available from the Twitter API through May, 2018; and (2) discussing a set of mining companies doing business in [Country], which were filtered by applying keyword query terms (e.g., `#adani, #santosltd`), and resulted in the tweet counts shown in Table 1.

Company	Instances
Adani	434057
Santos	92946
BHP	71473
RioTinto	29753
Woodside	17774
Fortescue	13779
Whitehaven	15205
Iluka	3243
OilSearch	2724
Cuesta	177
NewMont	4119

Table 1: Tweet counts by company

These raw tweets are pre-processed as follows.

- Tokenize texts using the CMU Tweet Tagger (Owoputi et al., 2013). Stop words (e.g., “not”) are retained.
- Remove “RT” tags marking re-tweets.
- Shrink character elongations (e.g., “yeees” → “yes”) except in usernames.
- Replace URLs, mentions, and year, time, cash and hashtag items with place holders (e.g., “slo_url”, “slo_mention”, . . .).
- Down-case all text.
- Remove tweets that are not labelled as some variant of English either by the Twitter or by Polyglot.[3]
- Remove the 49 tweets found to not be associated with any company.

The resulting, processed dataset has 676,349 items.

Testing Set: To build a gold-standard testing set, we sampled and manually labeled a set of 200

[2] E.g., it includes tweet instances that quote Christopher Hitchens, a well-known athiest, as stance-against atheism, e.g., “Morality is not derived from religion, it precedes it.”.

[3] Twitter mislabelled some English tweets as non-English. There were 28,821 tweets in the raw dataset marked as non-English by Twitter, of which only 5,581 were also classified as non-English by Polyglot.

tweet instances that included randomly sampled tweets from the most common companies: Adani: 50; BHP: 50; Santos: 50; a sampling of RioTinto/Fortescue tweets: 50. The tweets for each company are semi-balanced in that we took random samples from the raw dataset but then manually replaced some of the tweets marked with the most populous codes with replacement tweets marked with less populous codes. The replacement tweets were manually collected and not included in the training set.

The testing set included the target company for each tweet and access to the full tweet, which includes the tweet author's profile text and the tweet's discourse context of preceding and responding tweets. Three coders were asked to manually code each example in the testing set based on instructions that generally followed Mohammad et al's instructions but additionally allowed scrutiny of the tweet's context, including the author's identify and profile description, and other tweets in the current tweet's discourse context. The author's identity and profile description can be used to determine the author's general views, and the tweet conversation history can also help by placing the tweet in the context of other tweets that are for and/or against the target.

The inter-coder Fleiss Kappa score (Fleiss, 1971; Antoine et al., 2014) on this dataset is 0.64, which is commonly considered to be "substantial agreement" (Landis and Koch, 1977). From the 200 originally-coded tweets, we chose the majority code, throwing out examples where the three coders chose three options and where examples were coded as not-applicable (i.e., *na*). This resulted in a gold standard testing set of 177 items with the following distribution: Adani: 49; Santos: 43; BHP: 38; RioTinto: 35; Fortescue: 12.

Training set: The training set is populated using hand-built, auto-coding rules based on hashtags and author usernames strongly associated with one particular stance with respect to the Adani mining company. The rules are as follows.

- *#stopadani* → *stance-against* — This is overwhelmingly the most common hashtag (28080 instances).
- *#goadani, #stopstopadani* → *stance-for* — These hashtags are less common (621 instances).
- known news sources, e.g., sevennews → *stance-neutral* — These known, mainstream news sources are assumed to be generally even-handed in their treatment of the issues and are thus coded as stance-neutral.

A script uses these rules to collect one training set of candidate tweets from the full dataset for Adani, the most commonly mentioned mining company. The results are restricted to a fixed 1-1-1 ratio between for-against-neutral tweets, which led to a training set size of 621-621-621. The auto-coding hashtags are removed from the training set before training.

To test the accuracy of the auto-coding, we randomly sampled 24 tweets from the resulting training set and asked three coders to manually code them as stance for, against or neutral. The coding had a Fleiss Kappa score of 0.71, again in the "substantial agreement" range, and the majority code from this manual coding matched the auto-coding in all cases. We took this as evidence that the auto-coder based on the simple rules listed above is making accurate codings.

4 Task Setup

This section describes the model, the training process, and the results.

4.1 Model

Following SemEval-2016's *SVM-ngrams* baseline model, we use a single SciKit-Learn Linear SVC classifier (Pedregosa et al., 2011) with these input features:

- *N-gram counts*: A vector is generated from each tweet, components of which are counts of n-grams in the tweet: 2–5 character n-grams; 1–3 word n-grams, computed using SciKit-Learn's CountVectorizer. When the author profile text is included, it is concatenated with the tweet text and vectorized.
- *target company name presence*: Because we don't consider tweets that don't mention the target, this is always set to 1.
- *word embeddings*: Embeddings for a concatenation of the tweet and user profile tokens are produced, based on FastText word embeddings (Bojanowski et al., 2017) built from the tweet and user profile texts from the full processed dataset.

	Adani				BHP	Santos	RioTinto	Fortescue	Comb.
	For	Against	Neutral	Comb.					
No-Profile									
Mean	0.19	0.64	0.45	0.43	0.37	0.33	0.54	**0.46**	0.43
Median	0.19	0.67	0.46	0.41	0.37	0.34	0.55	0.47	0.43
Std.Dev.	0.11	0.09	0.029	0.05	0.05	0.07	0.06	0.10	0.03
With Profile									
Mean	**0.82**	**0.66**	**0.75**	**0.74**	**0.60**	**0.52**	**0.67**	0.45	**0.60**
Median	0.82	0.70	0.76	0.76	0.61	0.52	0.66	0.42	0.60
Std.Dev.	0.03	0.11	0.04	0.04	0.04	0.03	0.09	0.12	0.04

Table 2: Results for 25 runs of the SVM on balanced training sets tested both with and without profile texts

For each training/testing run, one SVM model is trained on an Adani-only training set and then tested on the gold-standard testing set, which contains tweets from Adani and from other companies.

4.2 Training

All training/testing runs were performed 25 times on separate, randomly-sampled (1-1-1) training sets and tested on the gold-standard testing set. Note that because stance-for is the rarest stance value for Adani, the 25 randomly sampled training sets would necessarily have the same stance-for tweets but would have potentially different, sampled tweets for the other stance codes. The average and standard deviation is computed for each set of training/testing runs.

The training is done both with and without profile texts, where each pair of runs uses the same base training set, with the profile texts either included as features or excluded.

4.3 Results

Table 2 shows the macro-F1 scores for 25X training/testing runs of the SVM with and without profile texts. The results for all stance codes are shown for Adani; only the combined scores for all codes are shown for the other companies. These results demonstrate two things.

First, adding author profile texts improves the performance of the classifier for all but one of the companies — see the improvement in the all-company-combined macro-F1 score from 0.43 to 0.60. This improvement occurred for all three code distributions: 1-1-1 (shown in the table) as well as the 1-5-1 and 1-10-1 distributions. The one exception to this result is Fortescue, which has the smallest number of tweet instances in the testing set and, not coincidentally, the highest standard deviation of the combined F1 score for any of the companies.

Second, the first four columns of the table show the relatively high performance on the Adani tweets of different stance codes and their combination, 0.74, compared with the combined scores for the other companies: BHP: 0.60; Santos: 0.52; RioTinto: 0.67; Fortescue: 0.45. The higher score for Adani is presumably due to the fact that the model was trained using an Adani-only training set; the lower but generally respectable scores for the other companies is presumably due to the common features shared by the different mining companies in the domain. Note that this result is not seen in the scores for the training without profile texts.

5 Conclusion

The results of the work described in this paper demonstrate that the use of an auto-coded training set is effective in the SLO domain and, presumably, in other SLO-related domains in which auto-coding rules can be derived by inspection of the data. Thus, identifying these auto-coding rules provides a way to obtain a training set that is of sufficient size and accuracy for stance classification. The results also demonstrate that the inclusion of the author's profile text as an input feature to the classifier improved the classifier performance. Thus, we expect that including these texts will improve the performance of stance classification in other SLO-related domains. Our future work includes broadening the strategy applied here to other domains apart from mining.

References

Jean-Yves Antoine, Jeanne Villaneau, and Anaïs Lefeuvre. 2014. Weighted Krippendorff's alpha is a more reliable metrics for multicoders ordinal annotations: Experimental studies on emotion, opinion and coreference annotation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 550–559. ACM. http://www.aclweb.org/anthology/E14-1058.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Neil Gunningham, Robert A. Kagan, and Dorothy Thornton. 2004. Social License and Environmental Protection: Why Businesses Go beyond Compliance. *Journal of the American Bar Foundation*, 39(2):307–341.

J. R. Landis and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33:159–174.

Saif M. Mohammad, Swetlana 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*, pages 31–41. ACM. https://www.aclweb.org/anthology/S/S16/S16-1003.pdf.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT 2013*, pages 380–390. ACL.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.