Moral Stance Recognition and Polarity Classification from Twitter and Elicited Text

Wesley Ramos dos Santos
University of São Paulo
São Paulo, Brazil
wesley.ramos.santos@usp.br

Ivandré Paraboni
University of São Paulo
São Paulo, Brazil
ivandre@usp.br

Abstract

We introduce a labelled corpus of stances about moral issues for the Brazilian Portuguese language, and present reference results for both the stance recognition and polarity classification tasks. The corpus is built from Twitter and further expanded with data elicited through crowd sourcing and labelled by their own authors. Put together, the corpus and reference results are expected to be taken as a baseline for further studies in the field of stance recognition and polarity classification from text.

1 Introduction

Computational sentiment analysis may be understood as a wide range of tasks intended to identify opinions, emotions and other types of stance expressed in natural language text (Tsytsarau and Palpanas, 2012; Liu, 2015). Among these, opinion mining is arguably the most well-studied form of sentiment analysis, consisting of identifying the target of an opinion, and/or the polarity (positive, negative, neutral etc.) of the sentiment expressed towards this target (Tsytsarau and Palpanas, 2012).

Stance recognition (Anand et al., 2011; Hasan and Ng, 2013; Lai et al., 2016; Mohammad et al., 2016b; Zarrella and Marsh, 2016; Wei et al., 2016; Mohammad et al., 2017), by contrast, consists of deciding whether the author of a piece of text shows a favourable or unfavourable attitude (or position) towards a certain target (Mohammad et al., 2017). The distinction between sentiment and stance is motivated by the observation that a sentiment, regardless of being positive or negative, may reflect a favourable or unfavourable position towards the target (Mohammad et al., 2016b). For instance, given the target topic 'veganism', a sentence as in 'beef tastes wonderful' expresses a positive feeling (which would indeed be recognised as such by traditional sentiment analysis systems), but it also reflects an unfavourable position towards this particular target.

Stance recognition from text is a well-known and yet challenging research topic. Systems of this kind enable the development of more complex sentiment analysis applications, and have been at the centre of a recent shared task (Mohammad et al., 2016b) focused on the use of supervised and unsupervised methods for stance recognition in the English language. For other less-resourced languages, however, resources remain scarce.

Based on these observations, this paper presents a labelled corpus of stances in Brazilian Portuguese, and a number of computational models addressing two related issues: *stance recognition*, is presently regarded as the binary classification problem of deciding whether a given text conveys *any* attitude towards a certain target topic or not, and *stance polarity classification*, which is regarded as the binary classification problem of deciding whether a given stance expressed as text shows a *positive or negative* attitude towards the target topic. Examples of both tasks for the target topic 'veganism' are as follows.

Stance recognition:

- *She says that avoiding animal products is just a fad (no stance towards veganism)*
- *Veganism will save the world! (a stance towards veganism)*

Stance polarity classification:

- *No one should ever eat beef (a positive stance towards veganism)*

Proceedings of Recent Advances in Natural Language Processing, pages 1069–1075,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_123

- *Vegans tend to have health issues (a negative stance towards veganism)*

As in the case of the English stance corpus in (Mohammad et al., 2016b), we will favour the recognition of stances about moral issues (e.g., abortion, drugs legislation etc.) in the Twitter domain. In addition to that, however, Twitter data will be presently expanded with a collection of moral stances elicited through crowd sourcing as well, and which were labelled by their own authors as a gold standard. Put together, the corpus and its reference results are expected to be taken as a baseline for further studies in moral stance recognition and stance polarity classification tasks.

2 Related Work

The work in (Anand et al., 2011) is among the first to address the computational recognition of stances from text, analysing a corpus of 4873 posts in on-line discussion forums. The data set considered covers 14 topics, ranging from entertainment to ideological issues. Favourable and unfavourable stances are recognised with accuracy of up to 69%, outperforming a unigram baseline model that obtained up to 60% accuracy.

Stance recognition in discussion forums is also addressed in (Hasan and Ng, 2013). In this case, however, the work focuses on the question of how the performance of a stance classifier varies in relation to the volume and quality of training data, regarding the complexity of the underlying model, the richness of the set of learning features and the use of extra-linguistic restrictions in a wide range of scenarios. The experiments leave a series of contributions on how to build models of this type, and on which kinds of knowledge to consider.

More recently, the SemEval-2016 competition (Mohammad et al., 2016a) brought together 19 participating systems engaged in the task of supervised stance recognition from tweets in the English language. The training corpus, described in detail in (Mohammad et al., 2016a), contains 2914 tweets about five target topics (atheism, climate change, feminism, Hillary Clinton and abortion legislation.) The corpus contains, on average, 583 tweets per target, but the set is unbalanced. On average, there are 25,8% positive and 47,9% negative stances. The test set, with 1249 tweets, is even more unbalanced, with 24,3% of positive stances and 57,3% negative stances. The SemEval corpus is the basis of some of studies discussed as follows.

The work in (Zarrella and Marsh, 2016) presents the best overall performance in the SemEval-2016 shared task (Mohammad et al., 2016b) on supervised stance recognition. The proposal makes use of a recurrent neural network with features learned by distant supervision from large unlabelled datasets. Word and phrase embedding models are trained using Word2Vec skip-gram (Mikolov et al., 2013), and then used for learning sentence representations with the aid of a hashtag prediction model. Finally, sentence vectors are optimised for stance recognition based on the labelled examples from the training corpus.

Also in the context of the SemEval-2016, the work in (Wei et al., 2016) presents an approach based on convolutional neural networks that, instead of simply predicting when the validation accuracy will reach its maximum, uses a voting scheme and other secondary improvements. The model is trained individually for each of the five targets of the SemEval-2016 corpus, and obtains the second best overall results for the supervised stance recognition track.

Subsequent to SemEval-2016, a number of improved systems have been proposed. The work in (Lai et al., 2016), for instance, explores the use of world knowledge - in the form of rules about friendships and political enmities - to enhance the task of recognising political stances in the SemEval-2016 corpus. The proposal consists of a stance recognition model enriched with semantic features of each target topic, which outperforms the participant systems of the original shared task.

Finally, the work in (Mohammad et al., 2017) presents a post-hoc evaluation of the SemEval-2016 stance recognition task, proposing a much simpler and more accurate model than the overall winner of the competition in (Zarrella and Marsh, 2016). The proposed model makes use of linear SVM and a set of features computed from the training data, such as word and character n-grams and word embeddings computed from an additional data set.

3 Current Work

The present investigation of moral stance recognition and polarity classification consists of a corpus data collection (described in Section 3.1), and two

individual experiments: stance recognition (Section 3.2) and stance polarity classification (Section 3.3). In both cases, we shall focus on methods that rely on lexical and morphological knowledge only by making use of word and char n-grams.

3.1 Corpora

Our initial goal was to create a corpus of moral stances in the Brazilian Portuguese language that would preferably be (a) at least as large as the English training dataset for SemEval-2016 supervised stance recognition task (Mohammad et al., 2016b), (b) more well-balanced if possible, and (c) not limited to the Twitter domain. To this end, we collected a 180k-word corpus conveying over 5,000 moral stances from two sources - Twitter and stances elicited through crowd sourcing - about five topics: abortion, death penalty, drug legalisation, criminal age, and racial quotas. Elicited texts are, on average, 3.5 times longer than tweets.

Corpus descriptive statistics for our two domains are summarised in Table 1.

Twitter messages were collected by searching Brazil Twitter for specific key words (e.g., 'abortion' etc.). For each topic, an initial 7000-message set was selected for manual inspection and labelling.

Elicited stances were obtained from a crowd sourcing task involving 490 Brazilian Portuguese native speakers. Participants were requested to give their opinions about each of the target topics by providing answers in a 0 (totally disagree) to 5 (totally agree) scale and, subsequently, were requested to provide motivation for each of their opinions by writing a short text. The elicited corpus has been subject to spell-checking and it is overall much more well-formed than the Twitter data, and with a larger vocabulary.

Twitter messages were manually labelled by assigning a positive/negative class to all messages that unequivocally expressed a stance on the intended topic, and by assigning the class 'other' to any message that did not meet these criteria. Thus, the class 'other' represents the fact that, despite containing a key word of interest, the message did not convey any obvious stance about the target topic, and it was therefore regarded as noise[1].

[1]For instance, annotators came across a number of references to 'Aborto Elétrico' (electrical abortion), which is the name of a rock band with no relation to any stance about the target topic.

Twitter text labelling proceeded until a minimum of 240 instances of each of the three class were identified, or until the end of the dataset was reached. This allowed us to obtain a certain balance between for/against stances for most topics, but resulted in a vast majority of samples labelled as 'other'. Thus, the 'other' class - which corresponds to non-stance text - is several times larger than the positive and negative classes in all five topics.

Elicited stances, by contrast, were assigned labels automatically based on the opinion scores provided by the crowd sourced participants. More specifically, scores 0 and 1 were taken as representing negative stances, 2 and 3 as neutral stances, and 4 and 5 as positive stances. Unlike the Twitter dataset, we notice that all elicited texts contain, by definition, some stance on the topics under discussion, and hence there is no 'other' (or non-stance) class in this domain.

Class label distributions for the Twitter and elicited datasets are summarised in Table 2.

3.2 Stance Recognition

Our first experiment - stance recognition - is presently defined as the binary classification problem of deciding whether a given text conveys any attitude towards a certain target topic or not. Since all texts from our elicited data (cf. the previous section) express, by definition, some stance about the target topic, the present task is applicable to Twitter data only.

3.2.1 Models

For the stance recognition task, a range of n-gram models - from 1 to 5 words and from 3 to 16 characters - was considered, and we found that character-based models always outperform word-based models. As a result, all models under consideration for this task are based on character n-grams only.

In what follows we consider the use of TF-IDF character counts (here by called our *Select.char* model) with k-best univariate feature selection using ANOVA F1 as a score function. By combining relatively long character sequences (which in most cases encompass words) with feature selection, we expect *Select.char* to outperform the alternatives under consideration, as discussed below.

The *Select.char* model was trained by making use of the best out of three possible learning methods - Naive Bayes, Logistic Regression and Mul-

Table 1: Corpus descriptive statistics

Source	vocab. size	words	messages	words / msgs
Twitter	5,789	44,564	2,792	16
Elicited	9,081	137,122	2,450	56
Overall	11,845	181,686	5,242	35

Table 2: Class distribution for Twitter and elicited data.

	Twitter stances			Elicited stances		
Topic	for	against	other	for	neutral	against
Abortion	240	384	2570	310	105	75
Death penalty	801	244	1518	105	125	260
Drugs	335	181	1482	263	129	98
Criminal age	243	240	1433	198	104	188
Racial quotas	240	364	2596	205	128	157
Overall	1859	1413	9599	1081	591	778

tilayer perceptron, with optimal k-values selected in the 5000 to 90000 range at 1000 intervals by performing grid search on the training dataset.

In addition to that, the entire input feature set (i.e., with no feature selection) is taken as the basis for two simpler methods - logistic regression (*LogReg.char*) and multilayer perceptron (*MLP.char*). The latter consists of 3 layers conveying 150 neurons each, and using rectified linear units (ReLU) as an activation function.

The three models of interest - *Select.char*, *LogReg.char* and *MLP.char* are to be evaluated against a majority class baseline *Majority*.

3.2.2 Data

The experiment makes use of the Twitter dataset described in Section 3.1 with random 80:20 train-test split.

3.2.3 Evaluation

Table 3 shows F1 results of stance recognition on Twitter data for both positive (stance) and negative (others, or non-stance) classes, and overall weighted F1 scores obtained by each model under consideration. Best weighted F1 scores for each target topic are highlighted.

As expected, all models easily outperform the *Majority* baseline, and the combination of char n-grams and feature selection in *Select.char* generally outperforms the alternatives under consideration, albeit for a small difference. This may be partially explained by the heavy data imbalance (cf. Table 2), which may have obscured possible differences across models. Moreover, we notice that variation across target topics is also small, suggesting that stance recognition is relatively topic-independent.

3.3 Stance Polarity Classification

Our second experiment - polarity classification - is presently defined as the binary classification problem of deciding whether a given stance expressed as text shows a positive or negative attitude towards the target topic. For this task we consider both elicited stances, and also the portion of Twitter data that conveys a positive or negative stance, that is, disregarding only those tweets labelled as 'other' (cf. section 3.1.)

3.3.1 Models

Given the overall positive results of character-based models and feature selection in the case of stance recognition (cf. the previous section), we will consider models of this kind for stance polarity classification as well. To this end, we make use of a char n-gram model - hereby called *MLP.char* that is similar to *Select.char* in the previous section, except that in the present case we will focus on the use of MLP classifiers only.

In addition to *MLP.char*, we also consider a mode based on skip-gram word embeddings (Mikolov et al., 2013), hereby called *MLP.w2vec*. The model makes use k-best univariate feature selection with the ANOVA F1 function over TFIDF-weighted word embeddings of size 50, 100 and 300. Learning methods under considerations are

Table 3: Weighted average F1 results for Twitter stance recognition

	Majority			LogReg.char			MLP.char			Select.char		
Topic	stance	other	avg	stance	other	avg	stance	other	avg	stance	other	avg
Abortion	0.00	0.89	0.71	0.40	0.85	0.76	0.43	0.88	**0.79**	0.42	0.89	**0.79**
Death penalty	0.74	0.00	0.44	0.62	0.74	0.69	0.65	0.78	0.73	0.72	0.81	**0.78**
Drugs	0.84	0.00	0.61	0.43	0.76	0.67	0.41	0.82	0.71	0.41	0.84	**0.72**
Criminal age	0.00	0.85	0.64	0.45	0.80	0.71	0.55	0.84	**0.76**	0.53	0.84	**0.76**
Racial quotas	0.00	0.89	0.70	0.36	0.85	0.75	0.36	0.86	0.76	0.33	0.89	**0.77**
Overall	0.32	0.53	0.62	0.45	0.80	0.72	0.48	0.84	0.75	0.48	0.85	**0.76**

MLP classifiers of 1-3 layers with numbers of neurons ranging from 33 up to the size of the embedding vector, and using either ReLU or hyperbolic tangent (tanh) as an activation function. Optimal parameters and vector sizes were determined by performing grid search on the training data.

Finally, given the affective nature of the topics in the corpus, we will also consider the use of psycholinguistic knowledge as provided by the LIWC dictionary (Pennebaker et al., 2001). LIWC models word categories such as love, money, power etc. that are known to play a significant role in a range of NLP tasks such as sentiment analysis and, in particular, personality recognition from text. Psycholinguistic knowledge will hence be the basis of a simple model - hereby called *LIWC* - consisting of a 64-feature subset of LIWC category counts for Brazilian Portuguese (Filho et al., 2013).

The three models of interest - *MLP.char*, *MLP.w2vec* and *LIWC* - are to be evaluated against two baseline systems: a majority class baseline *Majority*, and a word-based TFIDF model with k-best feature selection - hereby called *LogReg.word*, in both cases making use of logistic regression.

3.3.2 Data

The experiment makes use of the elicited stance dataset, and also the stance portion of the Twitter dataset as described in Section 3.1. In both cases, a random 80:20 train-test split was performed.

3.3.3 Evaluation

Table 4 shows weighted F1 score results obtained by each model under consideration for the Twitter domain, and Table 5 shows results for the elicited data. In both cases, best weighted F1 scores for each target topic are highlighted.

Although all Twitter models outperform the *Majority* baseline, results are overall mixed and, for two topics (death penalty and criminal age), the word-based baseline model *LogReg.word* actually outperforms the alternatives. Moreover, we notice that the psycholinguistics-based *LIWC* approach produces the second lowest results of all, and that none of the top-performing models seems clearly superior to the others. We hypothesise that the close results obtained by *LogReg.word*, *MLP.w2vec* and *MLP.char* may be partially explained by the use of the same underlying feature selection method, which turned out to be more significant than the actual choice of text representation or learning method.

Contrary to the Twitter scenario, results for the elicited data were uniform, with the *MLP.char* approach outperforming all alternatives by a large margin, and once again leaving the *Majority* baseline and *LIWC* models at the bottom. We hypothesise that, as in the case of the stance recognition experiment in the previous Section 3.2, the use of long char n-gram sequences does help the present task as well, and that it may have been particularly successful in combination with the higher text quality of elicited stances in our data.

Finally, a note on the use of char n-grams. As expected, the k-best n-grams in the models that use feature selection largely correspond to single words (e.g., 'unacceptable') or short expressions (e.g., 'I agree'), both of which clearly denoting stances in our domains. As a result, the model is comparable to a variable-length word n-grams, but with greater flexibility to include subwords (e.g., 'believ*'). To illustrate this, Figure 1 shows the char n-grams distribution among the k-best terms in the polarity classification task from the elicited dataset. From these results, we notice that the selected char n-grams largely fall within the 4..10 range, peaking at n-grams with a length of 6.

Table 4: Weighted average F1 results for polarity classification from Twitter data

Topic	Majority	LogReg.word	LIWC	MLP.w2vec	MLP.char
Abortion	0.41	0.68	0.61	0.66	**0.71**
Death penalty	0.53	**0.82**	0.67	0.81	0.77
Drugs	0.55	0.60	0.60	**0.77**	0.66
Criminal age	0.29	**0.82**	0.71	0.75	0.76
Racial quotas	0.49	0.73	0.70	**0.76**	**0.76**
Overall	0.45	0.73	0.66	**0.75**	0.73

Table 5: Weighted average F1 results for polarity classification from elicited data

Topic	Majority	LogReg.word	LIWC	MLP.w2vec	MLP.char
Abortion	0.49	0.69	0.52	0.76	**0.92**
Death penalty	0.37	0.53	0.43	0.72	**0.90**
Drugs	0.37	0.48	0.50	0.57	**0.82**
Criminal age	0.23	0.64	0.47	0.67	**0.87**
Racial quotas	0.25	0.50	0.42	0.59	**0.77**
Overall	0.34	0.57	0.47	0.66	**0.84**

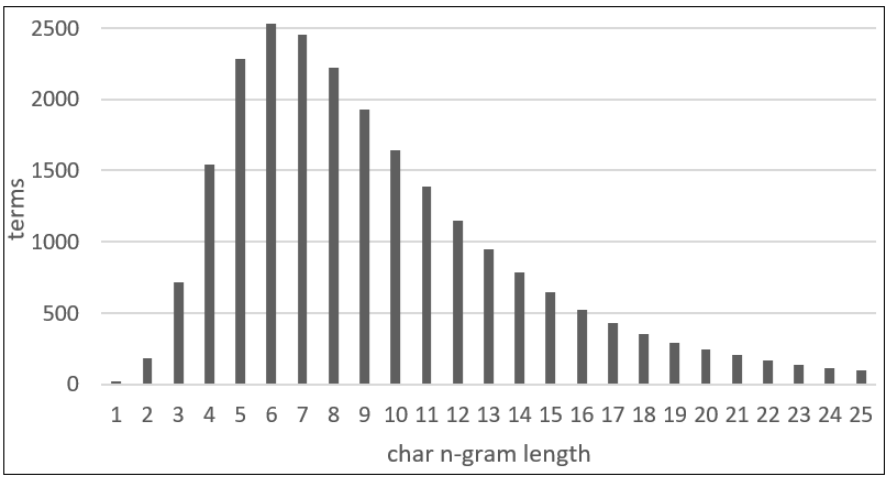

Figure 1: Char n-gram distribution across k-best terms in polarity classification from elicited data.

4 Final Remarks

This paper addressed the issue of moral stance recognition from text. We introduced a labelled corpus of stances taken from Twitter and additional crowd-sourced texts, and a number of supervised models of stance recognition and stance polarity classification.

Initial results suggest that both tasks may be performed with relatively high accuracy by making use of simple models based on char n-grams and feature selection. As expected, best results were observed when using more well-formed (in our case, crowd-sourced) texts, rather than when using Twitter data.

The corpus and the present results are expected to be taken as a reference for further studies in moral stance recognition in Brazilian Portuguese natural language processing. As future work, we intend to expand the current dataset in both domains by adding more instances and topics, and assess the use of deep learning methods for both the stance recognition and the polarity classification tasks.

Acknowledgements

This work received support by FAPESP grants # 2017/06828-1 and 2016/14223-0, and from the University of São Paulo.

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 (ACL-HLT 2011)*. Association for Computational Linguistics, Portland, Oregon, USA, pages 1–9.

Pedro P. Balage Filho, Sandra M. Aluísio, and T.A.S. Pardo. 2013. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. In *9th Brazilian Symposium in Information and Human Language Technology - STIL*. Fortaleza, Brazil, pages 215–219.

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

Mirko Lai, Delia Irazu Hernandez Farias, Viviana Patti, and Paolo Rosso. 2016. Friends and Enemies of Clinton and Trump: Using Context for Detecting Stance in Political Tweets. In *Mexican International Conference on Artificial Intelligence (MICAI-2016). Lecture Notes in Computer Science, vol 10061*. Springer.

Bing Liu. 2015. *Sentiment analysis: mining opinions, sentiments, and emotions*. Cambridge University Press.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, Curran Associates, Inc., pages 3111–3119.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. A dataset for detecting stance in tweets. In *Proceedings of 10th edition of the the Language Resources and Evaluation Conference (LREC-2016)*. Portoroz, Slovenia.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016b. Semeval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the International Workshop on Semantic Evaluation*. San Diego, California,USA.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *Special Section of the ACM Transactions on Internet Technology on Argumentation in Social Media* 17(3).

J. W. Pennebaker, M. E. Francis, and R. J. Booth. 2001. *Inquiry and Word Count: LIWC*. Lawrence Erlbaum, Mahwah, NJ.

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

Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. 2016. pkudblab at SemEval-2016 Task 6: A Specific Convolutional Neural Network System for Effective Stance Detection. In *Proceedings of the International Workshop on Semantic Evaluation*. San Diego, California,USA.

Guido Zarrella and Amy Marsh. 2016. MITRE at SemEval-2016 Task 6: Transfer Learning for Stance Detection. In *Proceedings of the International Workshop on Semantic Evaluation*. San Diego, California,USA.