Neural, Non-neural and Hybrid Stance Detection in Tweets on Catalan Independence

Wojatzki, Michael; Zesch, Torsten

DOI: https://doi.org/10.17185/duepublico/46448

URN: urn:nbn:de:hbz:464-20180627-144710-9

Link: https://duepublico.uni-duisburg-essen.de:443/servlets/DocumentServlet?id=46448

Source: Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017), Murcia, Spain, September 19, 2017, pp. 178-184

Neural, Non-neural and Hybrid Stance Detection in Tweets on Catalan Independence

Michael Wojatzki and Torsten Zesch

Language Technology Lab
University of Duisburg-Essen
Duisburg, Germany
michael.wojatzki@uni-due.de
torsten.zesch@uni-due.de

Abstract. We present our system LTL_UNI_DUE which participated in the shared task on automated stance detection in tweets on Catalan independence at IberEval 2017. In our system, we combine neural (LSTM) and non-neural (SVM) classifiers to a hybrid approach using a decision tree and heuristics.

1 Introduction

Recent political events have shown that political surveys often fail to predict the real outcome of elections. Examples of this miss-prediction could be observed in the vote on the UK exit from the European Union (*Brexit*) or the American presidential election. As one reason for this failure, it has been discussed that people behave in a socially desirable[1] manner in polling situations [6,10]. This effect could be circumvented by (additionally) examining data in which people naturally express their stances towards targets of interest. An obvious source for this data is social media, as stance taking is an essential part of social media interactions.

In order to reliably and efficiently conduct analyzes of such data, systems are needed that can automatically determine stance. To this end, NLP researchers have recently begun to systematically address social media stance detection. There were shared tasks on social media stance detection in English [8] and Chinese [13]. IBEREVAL2017 represents the first attempt to address this important task by providing data containing stance towards the target *Independence of Catalonia* [11]. During the training phase, the organisers released 4319 Tweets in Spanish and 4319 Tweets Catalan which were labeled with described the SemEval scheme. Participants could use this data to train stance detection systems that are subsequently evaluated on unknown test instances.

In the following, we describe our submission named LTL_UNI_DUE to this shared task. For our participation, we rely on the findings of previous shared tasks and develop a system that uses (almost) no language-specific models or tools and no additional training data.

[1] Behaving in a way that is more likely to have social approval.

2 System Description

Systems in the previous shared tasks SemEval 2016 task 6 [8] or the NLPCC Task 4 [13] are all based on supervised machine learning but show a significant variety. We reviewed the used approaches and could identify two major strands, namely neural architectures and more traditional classifiers. The first strand translates the training data in sequences of pre-trained word embeddings and feed these sequences into neural networks with Long Short-Term Memory (LSTM) or convolutional layers (cf. the two best team submissions in SemEval [14,12]). The second strand contains approaches which represent the data mostly through word and character ngrams, averaged word-embeddings and sentiment features (see [8,13]). These representations are subsequently used to train models with more traditional algorithms such as SVMs. The results of both shared task show that the second strand of classifiers is superior, but that the neural systems are highly competitive. For the participating system, we strive to combine the strengths of both strands. Consequently, we first implement prototypical representatives of the strands namely a neural architecture with a bidirectional LSTM layer in its core and an SVM.

Since both approaches require tokenized texts, we apply the Twitter-specific ArkTokenizer [4] from the DKPro Core framework (v1.9.0) [2] beforehand. In the present shared task, the organizers provide 4319 Tweets in Catalan and 4319 Tweets in Spanish, which can be used for training. We train a model for each of the provided languages separately, as it is unlikely that the lexicalized models strongly generalize across the languages. As this is – especially due to the close relationship of the languages – possible, future work should to examine this more closely.

2.1 Neural System

We implement a (bi-)LSTM neural network [9] using the Keras framework with the Theano backend[2]. The hyperparameters have initially been set based on literature, but have been iteratively optimized according to the training data and theoretical considerations. We otimized the hyperparameters by performing 10-fold cross-validation and tuning towards the highest micro averaged F_1-score. As our goal was to train a robust system, we chose the same hyper-parameters for which we reached an optimum in both languages.

As input we translate the training data into sequences of dense word vectors using the pre-trained vectors in Catalan and Spanish provided by [1]. The used word vectors were created by a model that extends the *skipgram* model by [7] with sub-word information and is thus expected to be more robust against morphological variations such as inflections.

The central bidirectional layer follows this layer and has *138* LSTM units, uses *tanh* activation and the *adam* optimizer [5]. Since we observe a divergence between the performance on train and test data over the epochs, we add a dropout of *0.2* between the forward and backward LSTM-layer and the embedding layer to enable regularization.

[2] `https://keras.io/`

Subsequently, we add another dense and a *softmax* classification layer. Due to the imbalance of the class distribution we train the network with sparse categorical cross-entropy as a loss-function. The network was trained *five* epochs with a batch size of 64.

2.2 Non-Neural System

The non-neural system is implemented using the DKPro-TC framework (v0.9.0) [3]. We represent the tweets as binary feature vectors of the top *3000* uni-, bi-, tri- word ngrams and bi-, tri-, and four- character ngrams. In addition, we add word embedding features derived from the above described pre-trained vectors [1]. For this purpose, we average the embeddings of all words in a tweet and add a feature per embedding dimension. Past shared tasks on stance detection demonstrate that it may be beneficial to utilize sentiment information e.g. from a sentiment lexicon. However, as we could not find a suitable and freely available sentiment detection tool or sentiment word list for both Spanish and Catalan, we don't utilize sentiment features. For training the model we rely on an SVM with a linear kernel provided by the DKPro-TC framework. Again, we tuned the hyperparameters by relying on a 10-fold cross-validation and the resulting micro averaged F_1-score.

2.3 Hybrid System

The consideration of the two strands of approaches in the past shared tasks has shown that they make different errors and also have different strengths. Consequently, the question arises whether one can combine the strengths of both models into a superior, hybrid system. To examine this question we built a third system that automatically decides whether a tweet should be classified with the neural or the non-neural system. Therefore, we first labeled every tweet with whether the SVM's respectively the LSTM's prediction was wrong or false.

We than train a classifier for each approach and each language to automate this decision. Since we want to base this decision on a simple set of rules which may be transferrable to other tasks, we use a decision tree for this classification. In detail, we use weka's J48 as implemented in DKPro-TC.

As features we use characteristics which are suspected of having an influence on the classifiability through the systems. We use the number of tokens per tweet as SVM and LSTM differ in the amount of context they model. As both systems are dependent on lexical redundancy between train and test data, we implement several redundancy features. These features are the type-token ratio and binary features indicating whether the tweet's n-grams are contained in the training data and whether there are a pre-trained embeddings for its tokens.

Table 1 shows the performance of this classification for both systems and both languages. The rather mediocre results leave huge room for future improvements and more sophisticated machine learning. Based on these classifications we conduct a final decision. In case the system could not derive preference towards one system as both systems are recommend or none, we rely on the SVM as its performance is overall better.

	Catalan	Spanish
SVM	0.75	0.66
LSTM	0.67	0.59

Table 1. Performance of the type prediction indicated by micro averaged F_1-score

3 Results on Training Data

In order to estimate the performance of our models, we evaluate them using a 10-fold cross-validation on the training data. Table 2 shows the performance of these experiments. For both the neural and non-neural approach, we observe better performance for Catalan than for Spanish. For the SVM we perceive an increase of $+0.1$ and for the LSTM we perceive an improvement of $+0.06$. Similarly, for both languages, the SVM performs significantly better than the LSTM. The performance decrease is bigger for Catalan (-0.1) than for Spanish (-0.06), which may be attributed to the overall better performance in Catalan.

We performed an ablation test at the level of feature groups to find out which data representation affects the model the most. The results are also shown in Table 2. We do not observe a large drop for any of the groups, which we attribute due to the fact that the modelled properties have strong overlap. For instance, embedding and unigram features model (almost) the same information, i.e. the occurrence of a certain word. However, unigrams are sparse and embeddings are dense word vectors, which both have specific advantages and disadvantages w.r.t. classification.

To quantify the similarity of the models we compute Cohen's κ, which is $\kappa = 0.28$ for Spanish and $\kappa = 0.39$ for Catalan. Since the predictions are clearly different, but both models show good performance, we conclude that there is in principle much room for the hybrid model. However, the hybrid system gains a performance similar to that of the SVM. When inspecting the similarity of the hybrid model and the SVM, we obtain $\kappa = 0.92$ (169 different predictions) for Catalan and $\kappa = 0.90$ (230 different predictions) for Spanish. This high degree of agreement between the models explains their similar performance. In order to demonstrate the upper bound of the hybrid system, we also compute a oracle condition in which we assume that the LSTM vs. SVM prediction was done correctly. This oracle condition results in an increase of performance of +0.09 for Catalan and + .13 for Spanish which demonstrates the potential of the approach.

In order to provide a deeper insight into the classification performance of the models, we show the corresponding confusion matrices in Table 3 for Catalan and in Table 4 for Spanish. For both languages, we observe for the SVM a more even error distribution than for the LSTM. However, the LSTM distributes its predictions mainly to the two frequent classes (FAVOR and NEUTRAL for Catalan and AGAINST and NEUTRAL for Spanish). The hybrid model combines these two tendencies by adjusting the prediction of the SVM towards the class distribution. However, thereby a similar proportion of advantageous and disadvantageous adjustments is made.

	Catalan	Spanish
SVM	0.80	0.70
- embeddings	0.79	0.70
- character ngrams	0.78	0.70
- word ngrams	0.78	0.68
(Bi-)LSTM	0.70	0.64
Hybrid	0.80	0.70
Oracle	0.89	0.83

Table 2. Micro averaged F_1-score obtained from the 10-fold cross-validation. For the SVM we show the results of an feature ablation test.

SVM

Actual \ *Predicted*	Against	Favor	Neutral
Against	44	42	45
Favor	33	2238	377
Neutral	38	336	1166

LSTM

Actual \ *Predicted*	Against	Favor	Neutral
Against	1	91	39
Favor	15	2085	548
Neutral	5	588	947

HYBRID

Actual \ *Predicted*	Against	Favor	Neutral
Against	42	47	42
Favor	29	2258	361
Neutral	30	369	1141

Table 3. Confusion matrices for Catalan

SVM

Actual \ *Predicted*	Against	Favor	Neutral
Against	967	60	419
Favor	97	97	141
Neutral	468	93	1977

LSTM

Actual \ *Predicted*	Against	Favor	Neutral
Against	669	41	736
Favor	104	25	206
Neutral	411	64	2063

HYBRID

Actual \ *Predicted*	Against	Favor	Neutral
Against	945	59	442
Favor	91	92	152
Neutral	461	94	1983

Table 4. Confusion matrices for Spanish

4 Results on Test Data

In this section we show, how the systems perform on the test data. We report the results in accordance with the official metric as defined by the organizers. The official metric is the macro-average of $F_1\oplus$ and $F_1\ominus$. Note that this metric is beneficial for systems which are similarly good at predicting $F_1\oplus$ and $F_1\ominus$, but punishes systems which are more imbalanced.

Table 5 gives an overview on the performance of our submission on the training data.

	Catalan	Spanish
SVM	0.43	0.42
(Bi-)LSTM	0.28	0.37
Hybrid	0.44	0.43

Table 5. Macro-average of $F_1\oplus$ and $F_1\ominus$ obtained from the train-test split [11].

Overall, we again observe that the SVM is superior to the LSTM system. The especially poor performance of the LSTM can also be explained by the used metric, which punishes the LSTMs tendency to ignore the sparse classes (FAVOR for the Spanish data and AGAINST for the Catalan data).

Similar to the results on the training data, we hardly see a difference between the hybrid and the SVM system for both languages. We attribute this again to the used heuristic, which uses the SVM prediction in cases were we cannot be sure about a decision. However, as the the hybrid and the SVM prediction is significantly different, we still see a high potential of hybrid approaches. As described above, future work should focus on a more accurate SVM or LSTM type prediction and more advanced heuristics.

5 Conclusion

In this work, we have described our participation in the shared task on automated stance detection in tweets on Catalan independence at IberEval 2017. The presented system relies on i) neural (LSTM) classifiers, ii) non-neural (SVM) classifiers and a hybrid approach which combines both classification paradigms on the basis of a decision tree and heuristics. On both the train and the test data we could not demonstrate a clear superiority of a hybrid approach. However, the obtained results highlight the potential of hybrid attempts and promising directions for further improvements.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group ”User-Centred Social Media”.

References

1. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
2. Eckart de Castilho, R., Gurevych, I.: A broad-coverage collection of portable nlp components for building shareable analysis pipelines. In: Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT. pp. 1–11. Dublin, Ireland (2014)
3. Daxenberger, J., Ferschke, O., Gurevych, I., Zesch, T., et al.: DKPro TC: A Java-based Framework for Supervised Learning Experiments on Textual Data. In: ACL (System Demonstrations). pp. 61–66. Baltimore, USA (2014)
4. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A.: Part-of-speech tagging for twitter: Annotation, features, and experiments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2. pp. 42–47. Portland, USA (2011)
5. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 pp. 1–13 (2014)
6. Krysan, M., Couper, M.P.: Race in the live and the virtual interview: Racial deference, social desirability, and activation effects in attitude surveys. Social psychology quarterly pp. 364–383 (2003)
7. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
8. Mohammad, S.M., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C.: Semeval-2016 task 6: Detecting stance in tweets. In: Proceedings of the International Workshop on Semantic Evaluation (to appear). San Diego, USA (2016)
9. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing 45(11), 2673–2681 (1997)
10. Streb, M.J., Burrell, B., Frederick, B., Genovese, M.A.: Social desirability effects and support for a female american president. Public Opinion Quarterly 72(1), 76–89 (2008)
11. Taulé, M., Martí, M.A., Rangel, F., Rosso, P., Bosco, C., Patti, V.: Overview of the task of stance and gender detection in tweets on catalan independence at ibereval 2017. In: Notebook Papers of 2nd SEPLN Workshop on Evaluation of Human Language Technologies for Iberian Languages (IBEREVAL). Murcia, Spain (2017)
12. Wei, W., Zhang, X., Liu, X., Chen, W., Wang, T.: pkudblab at semeval-2016 task 6: A specific convolutional neural network system for effective stance detection. In: Proceedings of the 16th International Workshop on Semantic Evaluation. pp. 384–388. San Diego, USA (2016)
13. Xu, R., Zhou, Y., Wu, D., Gui, L., Du, J., Xue, Y.: Overview of NLPCC Shared Task 4: Stance Detection in Chinese Microblogs. In: International Conference on Computer Processing of Oriental Languages. pp. 907–916. Springer (2016)
14. Zarrella, G., Marsh, A.: MITRE at semeval-2016 task 6: Transfer learning for stance detection. In: Proceedings of the 16th International Workshop on Semantic Evaluation. pp. 458–463. San Diego, USA (2016)