EmpNa at WASSA 2021: A Lightweight Model for the Prediction of Empathy, Distress and Emotions from Reactions to News Stories

Giuseppe Vettigli
Centrica plc
giuseppe.vettigli@centrica.com

Antonio Sorgente
Institute of Applied Sciences
and Intelligent Systems
National Research Council
antonio.sorgente@isasi.cnr.it

Abstract

This paper describes our submission for the WASSA 2021 shared task regarding the prediction of empathy, distress and emotions from news stories. The solution is based on combining the frequency of words, lexicon-based information, demographics of the annotators and personality of the annotators into a linear model. The prediction of empathy and distress is performed using Linear Regression while the prediction of emotions is performed using Logistic Regression. Both tasks are performed using the same features. Our models rank 4th for the prediction of emotions and 2nd for the prediction of empathy and distress. These results are particularly interesting when considered that the computational requirements of the solution are minimal.

1 Introduction

In recent years the NLP community has put particular effort into the identification of emotions in natural language. Methodologies based on Deep Learning are driving these efforts as they are currently top-performing on all the tasks proposed until now. This is reflected by the results of the shared tasks at the recent editions of WASSA (Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis). The shared task of WASSA 2017 (Mohammad and Bravo-Marquez, 2017) proposed to automatically identify the intensity of anger, fear, joy, and sadness from tweets. The task saw 21 participants and the winner used an ensemble model that combines three different types of Deep Neural Networks (Goel et al., 2017).

The Shared task EmoInt, hosted at WASSA 2018 (Klinger et al., 2018) proposed to identify emotions from tweets where words denoting emotions were removed. The winning solution (among 30 submissions) used an ensemble model consisting of language models together with a LSTM-based network containing a Convolutional Neural Network (CNN) as attention mechanism (Rozental et al., 2018).

The shared task at WASSA 2021 proposes to evaluate levels of empathy and distress and predict the emotion from news stories (Tafreshi et al., 2021). This new element around empathy and distress is of particular interest as it provides the community with a dataset where the annotations are supported by proper psychological theories regarding empathy and distress.

The same authors of the dataset have already shown that Deep Learning approaches can beat simpler methodologies, such as Ridge Regression, for the prediction of empathy and distress in (Buechel et al., 2018). More recently, a Mixed-Level Feed Forward Neural Network was also proven to be effective for the creation of lexicons for empathy and distress (Sedoc et al., 2020). Structures based on LSTM and CNN have also been proven to be effective on a separate dataset (Khanpour et al., 2017).

In a scenario where Deep Learning techniques have top performances but are also very demanding in regards to computational resources, we want to challenge the status quo by proposing a shallow model based on linear regression that has minimal computational requirements for both the tasks proposed at WASSA 2021. Thanks to specifically handcrafted features, the models that we propose have results comparable to much more sophisticated solutions already found in literature and, notably, can be trained on commodity hardware within seconds.

This paper is organized as follows. In Section 2 we will briefly introduce the dataset and the task. In Section 3 we will introduce the model used for the prediction. In Section 4 we will discuss the results. Finally, in Section 5 we will offer some conclusions regards our work and the directions of

April 19, 2021. ©2021 Association for Computational Linguistics

our future efforts.

2 Data and Task

The data includes 2130 essays, 270 released in a development set and 1860 released in the train set, annotated with an index between 1 and 7 for empathy and distress. Each essay is also associated with an emotion label among sadness, anger, neutral, fear, surprise, disgust, joy.

The task challenges the participant in predicting empathy, distress and emotion.

The essays are reactions to news articles and have length between 280 and 800 characters. This is an example of an essay from the dataset:

> “it is frightening to learn about all these shark attacks but these surfers should be aware of the risks associated with the sport. relocating the sharks should be a priority and it would be in the best interest to establish a moratorium on water sports until the shark epidemic is dealt with. closing beach in australia is a good precautionary method”

This essay is annotated with an empathy of 4.167 and a distress of 5.250 while the emotion is fear.

One of the most interesting features of this dataset is that, for each essay, it reports the following attributes regards the annotators: gender, education, race, age, income, personality conscientiousness, personality openness, personality extraversion, personality agreeableness, personality stability, interpersonal perspective taking, interpersonal personal distress, interpersonal fantasy, interpersonal empathetic concern.

The submissions to the challenge are evaluated on a test set of 525 samples. At the moment that we are writing the labels for the test set have not been released.

A restricted version of this dataset was initially introduced in (Buechel et al., 2018).

3 A Unified Model

The idea behind our approach is to achieve competitive results using well known tools that can be used on commodity hardware.

We build the features representing the text as n-grams and adding a set of characteristics extracted from a handcrafted set of lexicons. We decide to use Linear Regression for the prediction of empathy and distress and Logistic Regression for the prediction of emotions.

For the extraction of the lexical features and the creation of the prediction models, we use the scikit-learn library (Pedregosa et al., 2011).

3.1 Features

We combine features of three different types. Lexical features, extracted solely from the essays. Features reflecting the demographic and personality of the annotator that come along with the original data. Lexicon-based features extracted from a set of well known lexicons historically useful for the identification of emotions and Sentiment Analysis.

Lexical. Before extracting the lexical features we remove words with high frequency to avoid stop words, then normalize the remaining words using the Porter Stemmer via the NLTK implementation (Bird et al., 2009). Finally, we extract a set of n-grams from each essay and represent them using *tf-idf* (Salton and McGill, 1986).

Demographics and Personality. All the features regarding the demographics and the personality of the annotators. These features were normalized using standard scaling.

Lexicon-based. We consider a set of lexicons annotated with different psychological aspects. Each lexicon contains words annotated with a binary label or an intensity value. For each lexicon, we compute the average score for each word in the essay (binary labels are considered as 1 or -1). All the scores are considered as input features of the model.

This is the list of lexicons considered:

- Opinion Lexicon, words annotated as positive or negative towards opinion or sentiment (Hu and Liu, 2004).
- AFINN, list of words rated for valence with an integer between minus five (negative) and plus five (positive)(Nielsen, 2011).
- General Inquirer lexicon, list words classified as positive or negative according to the psychological Harvard-IV dictionary (Stone et al., 1966).
- Sentiment140 Lexicon, list of words annotated with a real-valued score from tweets with emoticons (Mohammad et al., 2013).

- +/-Effect Lexicon, list of words annotated as positive or negative with respect to the opinions expressed toward the effect that events have on entities (Choi and Wiebe, 2014).

- QWN, words annotated as positive or negative using the Q-WordNet PPV method (San Vicente et al., 2014).

- Twitter, list of words annotated using label propagation using the method described in (Speriosu et al., 2011).

- SenticNet, lexicon that contains a collection of concepts annotated with different values, in this work we consider only polarity, temper, attitude and sensitivity (Cambria et al., 2010).

- Affective rating, list of words annotated with affective norms (Warriner et al., 2013).

The scores extracted from the lexicons are also normalized using standard scaling.

3.2 Prediction of Empathy and Distress

The prediction of empathy and distress is done by combining all the features into a linear model. More specifically, the target variable is predicted as

$$\hat{y} = w^{(T)}x^{(T)} + w^{(D)}x^{(D)} + w^{(L)}x^{(L)} + w_0, \quad (1)$$

where $x^{(T)}$ contains the *tf-idf* representation of all the terms considered, $x^{(D)}$ the demographic and personality features and $x^{(L)}$ contains the features derived from the lexicons. The vectors $w^{(\cdot)}$ are weights to estimate while w_0 is a scalar representing a bias term. The system is trained computing the block vector of weights $\mathbf{w} = (w^{(T)}, w^{(D)}, w^{(L)}, w_0)$ as

$$\mathbf{w} = \arg\min_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2,$$

where $\mathbf{X}$ is a matrix where each row contains all the features extracted from a given essay, plus a unit value to take into account the bias weight w_0, and $\mathbf{y}$ is the vector of targets. More formally each row of $\mathbf{X}$ is defined as $\mathbf{x} = (x^{(T)}, x^{(D)}, x^{(L)}, 1)$. Note that $\mathbf{w} \in \mathbb{R}^{F+1}$ and $\mathbf{X} \in \mathbb{R}^{N \times F+1}$ where F is the total number of features and N the number of essays available.

The training is done separately for empathy and distress.

The lexical features consider uni-grams, bi-grams and tri-grams. During pre-processing 20% of the most frequent words were removed for the model that predicts empathy and 30% of the most frequent words were removed for the model that predicts distress. These parameters have been selected using a Grid Search.

3.3 Prediction of Emotions

For the prediction of emotions we extend the model presented above to compute the probability of the essay presenting a specific emotion. We compute the probability of an emotion e as

$$P_e = \frac{1}{1 + exp(-\gamma)},$$

where γ is given by the linear combination in equation 1. The emotion related to the essay is predicted as the emotion with the highest probability.

The system this time is trained finding the vector $\mathbf{w}$ as

$$\mathbf{w} = \arg\min_{\mathbf{w}} \sum_{i=1}^{N} log(1 + e^{-y_i \mathbf{x}_i \mathbf{w}^\top}) + \lambda \|\mathbf{w}\|_2^2$$

Where λ is a multiplier that allows us to regularize the model.

For this model we remove the 10% most frequent words and consider only uni-grams and bi-grams. We also truncate the number of terms extracted to 1000. The regularization parameter is set to $\lambda = \frac{1}{0.9}$. These parameters have also been selected using a Grid Search.

4 Results

To build the model for the identification of empathy and distress we used 10-fold cross-validation for an initial evaluation and comparison with the results available in the literature. We considered the Pearson correlation, metric adopted for the competition, as the main score but we also considered the Mean Absolute Error (MAE) and the Mean Squared Error (MSE).

In Table 1 we compare our model to the known state of the art, prior to the competition, given by the CNN-based solution introduced in (Buechel et al., 2018). In this comparison, it is important to keep in mind that the CNN model only uses textual data.

In the Table we note that our model restricted to n-grams has results that are already comparable to

target	features	Pearson	MAE	MSE
empathy	CNN from (Buechel et al., 2018)	0.404	-	-
empathy	only n-grams	0.390	2.908	1.413
empathy	n-grams + lexicons	0.460	2.623	1.320
empathy	n-grams + demographics/personality	0.493	2.596	1.322
empathy	all	**0.496**	2.606	1.335
distress	CNN from (Buechel et al., 2018)	0.444	-	-
distress	n-grams	0.424	2.529	1.307
distress	n-grams + lexicons	**0.512**	2.508	1.307
distress	n-grams + demographics/personality	0.494	2.491	1.294
distress	all	0.501	2.511	1.291

Table 1: Performances of the model with different set of features estimated using 10-fold cross-validation.

the CNN and that the CNN is outperformed when adding more features. We also note that demographics and personality features are particularly effective in predicting empathy while the lexicon-based features are able to give a major boost for the prediction of distress.

Considering the result on the final test set of the competition our model achieves a Pearson correlation of 0.516 for the prediction of empathy and 0.554 for the prediction of distress. The final score for the competition, computed averaging the results of both the predictions, is 0.554 and ranks 2nd in the final leader board with a difference of 0.009 from the highest score.

The results of the model for the prediction of emotions were less encouraging as we estimated a macro F1-Score of 0.330 and micro F1-Score of 0.413 using 10-fold cross-validation. While on the final test set the model achieves a Macro F1-Score of 0.313 and a Micro F1-Score of 0.396. These results are placed at the 4th position. Considering that the winning model achieved a macro F1-Score of 0.553 and a micro F1-Score of 0.623 we can conclude that our model is not competitive in the prediction of emotions.

5 Conclusions

Our submission demonstrated the effectiveness of using a shallow model with carefully handcrafted features for the prediction of empathy and distress. For this task we were able to beat the state of the art, previous to the competition, and achieve a highly ranked position on the leader board. An important result of our work is that our model demands only a fraction of the computational resources compared to the other models available.

We also showed that our model is too simplistic for the prediction of emotions and more sophisticated approaches are necessary to achieve good results.

From studying our models we realized that they have limited elements to be explained and we would like to improve this in our future efforts. In particular, we would like to study which lexical patterns and features drive the predictions of empathy and distress.

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765, Brussels, Belgium. Association for Computational Linguistics.

Erik Cambria, Robert Speer, Catherine Havasi, and Amir Hussain. 2010. Senticnet: A publicly available semantic resource for opinion mining. In *AAAI fall symposium: commonsense knowledge*, volume 10. Citeseer.

Yoonjung Choi and Janyce Wiebe. 2014. +/-effectwordnet: Sense-level lexicon acquisition for opinion inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1181–1191.

Pranav Goel, Devang Kulshreshtha, Prayas Jain, and Kaushal Kumar Shukla. 2017. Prayas at emoint 2017: An ensemble of deep neural architectures for emotion intensity prediction in tweets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 58–65.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. 2017. Identifying empathetic messages in online health communities. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 246–251.

Roman Klinger, Orphee De Clercq, Saif Mohammad, and Alexandra Balahur. 2018. Iest: Wassa-2018 implicit emotions shared task. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 31–42.

Saif Mohammad and Felipe Bravo-Marquez. 2017. Wassa-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 34–49.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrccanada: Building the state-of-the-art in sentiment analysis of tweets. In *In Proceedings of the International Workshop on Semantic Evaluation, SemEval '13*.

Finn Årup Nielsen. 2011. A new anew: Evaluation of a word list for sentiment analysis in microblogs. In *Workshop on'Making Sense of Microposts: Big things come in small packages*, pages 93–98.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Alon Rozental, Daniel Fleischer, and Zohar Kelrich. 2018. Amobee at iest 2018: Transfer learning from language models. *arXiv preprint arXiv:1808.08782*.

Gerard Salton and Michael J McGill. 1986. Introduction to modern information retrieval.

Inaki San Vicente, Rodrigo Agerri, and German Rigau. 2014. Simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 88–97.

João Sedoc, Sven Buechel, Yehonathan Nachmany, Anneke Buffone, and Lyle Ungar. 2020. Learning word ratings for empathy and distress from document-level user responses. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1664–1673.

Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge. 2011. Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 53–63.

Philip J Stone, Dexter C Dunphy, and Marshall S Smith. 1966. The general inquirer: A computer approach to content analysis.

Shabnam Tafreshi, Orphée De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, and Alexandra Balahur. 2021. WASSA2021 Shared Task: Predicting Empathy and Emotion in Reaction to News Stories. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.