Text Classification for Subjective Phenomena on Disaggregated Data and Rater Behaviour

Ewelina Gajewska[1], Barbara Konat[1]

[1]Faculty of Psychology and Cognitive Sciences,
Adam Mickiewicz University, Poznań, Poland
ewegaj@st.amu.edu.pl, bkonat@amu.edu.pl

Abstract

Phenomena such as emotional experience and offensive language perception are highly subjective in nature. Yet, the dominant approach in building automatic emotion and hate speech detection systems is based on the opinion of the majority. Recently, however, a personalised or human-centred approach has been proposed by the computational social scientists. In the current paper, we propose a novel method for modelling individual perspective in emotion detection and abusive language recognition, following existing works in this area (Miłkowski et al., 2021). We show that the personalised approach that implements our *Personalisation Metric* (PM) outperforms traditional majority-based methods in regard to subjective phenomena such as emotion and abusive language detection. Proposed method could be successfully used in the development of more accurate classification models suitable for the opinions of individuals as well as in recommendation systems.

Keywords: EDO 2023, emotion recognition, human-centred NLP, offensive language, recommendation systems

1. Introduction

Emotion detection in textual data has become a topic of interest in recent years, both for academic scholars and the industry. Emotion analysis is commonly employed in mining customer opinions and investigating public attitudes towards candidates in political campaigns. It is also used in psychotherapy sessions for assessing emotional states of patients (Tanana et al., 2021).

On the other hand, the growth of social media foregrounded the problem of hate speech and offensive language. However, currently available tools are insufficient to deal with the moderation problem. Thus, offensive language detection remains challenging due to the subjective nature of the task and the quality of data annotation, which is usually based on the opinion of the majority (Binns et al., 2017; Sap et al., 2021). For example, Waseem (2016) found systematic differences in the annotations given by the experts (anti-racism activists) and the crowd on hate speech. Therefore, the aggregated or the average vote does not fully represent the viewpoint of any side here and cannot be regarded as high quality data. Moreover, Sap et al. (2021) investigated demographic and political factors that influence the ratings on toxic language. The results indicate that more conservative annotators are more likely to rate African American English dialect as toxic, for example.

In recent years, researchers proposed to change the perspective from data-centric to human-centred in the NLP field, accounting for social factors in tasks such as offensive language detection, sentiment analysis and sarcasm recognition (Kocoń et al., 2021). However, techniques proposed so far are limited to specific types of data (numeric label scores in the case of (Miłkowski et al., 2021)) or require additional information about individuals such as demographic features, personality types or previous activity on social media platforms that could be difficult to obtain (Hovy, 2015; Lukin, Anand, Walker and Whittaker, 2017; Kocoń et al., 2021). The proposed Personalisation Metric (described in detail in Section 3) does not rely on any private information about users and could be calculated based only on labels given by individuals. Moreover, it is designed for categorical labels, which are predominant in machine learning tasks.

Researchers in the field also called for releasing annotator-level data instead of only so-called "ground truth" labels obtained through the majority voting technique (Prabhakaran, Davani and Diaz, 2021). Some noticed that those aggregated labels do not take into account perspectives of minority groups and that they do not even reflect the general opinion of the "average Jane" (Rizos and Schuller, 2020). Akhtar, Basile and Patti (2021) state that the NLP field needs novel methods to model conflicting perspectives in the automated systems for abusive language detection. Instead of majority-voted labels, they propose to differentiate different groups of individuals based on socio-demographic factors and develop separate gold standards and algorithms for each of them. Similarly, Basile (2020) emphasises that current machine learning techniques in NLP need to be adapted to subjective and pragmatic phenomena in order to create fair and inclusive models.

2. Related Work

In the current paper we adopt a human-centred approach introduced in (Kocoń et al., 2021). In particular, we focus on the microscopic level – the perspective of individuals on highly subjective tasks such as emotion and abusive language perception. Kocoń et al. (2021) differentiate also the mesoscopic and macroscopic levels, that is the perspective of selected groups of individuals and the general view, respectively. In their own study models

developed on the group-based labels performed better than the general view approach. For the microscopic level they include ids of annotators that also boost the performance of offensive content classifiers.

Previous works show also the improvement of classification performance on sentiment analysis and topic identification when demographic factors are taken into account (Hovy, 2015).

Several works reported high disagreements between raters on hate speech annotation and therefore called for modelling annotator perspectives in hate speech detection systems (Binns et al., 2017; Akhtar, Basile and Patti, 2020). Moreover, Larimore et al. (2021) demonstrate that racial identity as well as textual features of tweets influence annotator perceptions of racism. Similar problems could be observed in sentiment annotation. For example, Prabhakaran, Davani and Diaz (2021) found that around one-third of annotators achieve very low agreement scores with the majority voted labels.

Davani, Díaz and Prabhakaran (2022) experiment with 3 different techniques to implement individual perspectives into machine learning models on detection of online abuse and emotions. First, they present a multi-task approach that treats the prediction of labels for each annotator as separate subtasks. Second, they conceptualise the task as a multi-label classification where each label corresponds to individual annotators' labels. Third, the authors train an ensemble of models, one dedicated for each annotator where a final label, however, is majority voted. Alternatively, they propose to estimate uncertainty of a model's prediction to reflect the disagreements in the annotation.

Another strand of work that shows the importance of a personalised approach is the study of annotator bias in data. Results of (Wich, Bauer and Groh, 2020) indicate that political bias negatively impacts the automated detection of hate speech and in result could lead to racial discrimination.

3. Methodology

3.1. Personalisation metric

Proposed in the current paper *Personalisation Metric* (PM) extends *Personal Emotion Bias* (PEB) measure introduced by Miłkowski et al. (2021). PEB is based on Z-scores between the average and the individual user rating and is suitable for real-valued ratings (numeric variables) of emotional intensity. We introduce the PM metric that is suitable for categorical labels instead. PM comprises two well-known measures of agreement, i.e. the Cohen's Kappa statistic (Cohen, 1960) and percent agreement (accuracy). Therefore, the PM metric measures the similarity of opinions between two entities given a set of categories. Here, we consider the majority as one entity and an individual rater as the other entity.

We use accuracy and Kappa are complementary statistics here because of the so-called prevalence problem (generated by an imbalanced distribution of categories) that might yield low Kappa coefficients in cases where accuracy indicates almost perfect agreement between two entities (Eugenio, Glass, 2004). As a result, each rater is assigned with two values – Kappa and accuracy score which we jointly call the PM metric. In order to calculate these values, a sample of data annotated by an individual is compared against the same sample of data annotated with majority-voted labels. Thus, the PM metric indicates how similar is the opinion of an individual rater compared to the majority of people.

In both statistics, the values close to 1 signal perfect agreement, whereas scores around 0 indicate no agreement in the case of accuracy, and random agreement in the case of Cohen's Kappa. Negative values of Kappa reflect less than random agreement between two entities. Thus, we treat disagreements as an additional source of information instead of noise, which is common in aggregation-based approaches.

3.2. Model

Although Transformer-based architectures are currently considered state-of-the-art, Convolution Neural Networks (CNN) as well as Recurrent Neural Networks (RNN) still achieve good performance and are commonly used in text classification tasks (Tam, Said, Tanriöver, 2021). In particular, models based on the combination of CNN and RNN networks achieve superior performance compared with other architectures (Wang, Jiang and Luo, 2016). Therefore, our model makes use of CNN and bidirectional Long Short Term Memory (BiLSTM) networks. As a text representation method we employ GloVe 100-dimensional embeddings (Pennington, Socher, Manning, 2014). ReLu is the activation function in all layers except the last classification layer where we use sigmoid or softmax function in the case of emotion and abuse detection, respectively. Summarised description of the model used in the study can be found in Table 1.

Layer	Parameters
Embedding	GloVe 100-dim
CNN	Filters: 128; size: 5
MaxPooling	Size: 3
BiLSTM	Units: 64
GlobalMaxPooling	–
Dense	Units: 264
Dropout	Rate: 0.3
Concatenate	–
Dense	Units: 128

Table 1. Architecture of the proposed CNN-BiLSTM model.

3.2.1. Emotion detection

We make use of the *GoEmotions* dataset (Demszky et al., 2020) collected from Reddit and annotated with 27 categories of emotion. Here, in addition to majority voted labels, the authors release all individual annotations assigned by different raters. It therefore allows us to study individual perspectives in regard to emotion perception. For the purpose of the study, we included annotations only on 6 basic emotions in Ekman's taxonomy – fear, anger, sadness, surprise, joy, disgust, as well as a neutral category. We also decided to discard raters that annotated less than 334 data points (25th percentile) in order to

achieve stable scores between train and test sets for the PM metric. In addition, in order to have more data available for the training we merged annotations on selected emotions into the chosen 6 categories based on the correlation analysis conducted by the authors of GoEmotions. In result, the available corpus comprises over 128k data points (over 50k unique comments) that were annotated by 61 raters in total.

The GoEmotions dataset allows for multi-category multi-label classification of emotions, as annotators were allowed to indicate more than one emotion in a given Reddit comment. The data is however highly imbalanced as the neutral label is present in 42% of texts, and other categories are observed in 4% to 22% of the cases.

We design two conditions with respect to emotion detection at the individual level. First, we calculate the PM metric between each annotator and the majority voted label, separately for the train and test sets. Separate calculation of PM values for users in a test set allows to imitate a new set of users instead of copying PM scores from the train set. Thus, training is conducted on a different set of users and PM scores than evaluation. PM scores comprise here additional features next to text embeddings (PM condition). Second, we use ids of annotators transformed into one-hot encoded vectors as a set of additional features following related works (Kocoń et al., 2021) (vector-id condition).

In addition, we compare the performance achieved by those models with the traditional approach to text classification, i.e. based on the opinion of the majority (majority condition). Here, we make use of the results reported by the authors of the GoEmotions (Demszky et al., 2020) – BERT model fine-tuned for the classification of Ekman's 6 basic emotions. Therefore, we could compare not only the performance of the proposed approach (a model with PM metric features) with another method designed for individualised emotion recognition, but also the usefulness of a personalised approach with the traditional one based on the majority aggregated labels.

3.2.2. Abusive language detection

With respect to abusive language detection, we use the subset of 4k examples of the *ConvAbuse* corpus released by the authors (Curry, Abercrombie and Rieser, 2021). It encompasses short human-machine dialogues. Each data sample comprises 4 dialogical turns – 2 generated by a chatbot (conversational AI system) and 2 created by a user. For the purpose of our study, we concatenate all 4 turns into one text for each example that is later fed to a deep learning model. We made use of the annotations on abusive language on a 5-point scale: non-abusive, ambiguous, negative and mildly offensive, negative and insulting/abusive attitude, and strongly negative with overt incitement to hatred, violence or discrimination. The authors also make available individual annotations assigned by human subjects which allows us to study the subjectivity of abusive language perception and develop classification models that account for the individual view on abusive content perception. In total, there are over 12k instances of text annotated by 8 raters. Similarly, as in the emotion detection task, the data is highly imbalanced – over 78% of cases are assigned with the ambiguous category, and the other classes comprise from 2% to 7% of the data.

In regard to abusive language detection, we compare two approaches to text classification – the personalised one, which makes use of the PM metric (PM condition) and the popular majority-based approach (majority condition). In addition, we conduct cross-examination, i.e. training a model on majority aggregated labels and testing on individual annotations (cross condition). Similarly as in the case of emotion detection, we compute the PM metric for each annotator, separately for a train set and a test set. The PM metric comprises a pair of additional features fed to a model, next to word embeddings. In regard to the majority and cross conditions, text features (GloVe embeddings) comprise an input to a model. We release our source code regarding the abusive language detection study in Google Colab[1].

4. Results

We report the average results from 5 random splits of data. Each time the training set comprises 80% of data, and the remaining 20% is used for evaluation purposes. Performance is evaluated in terms of macro-averaged F1 scores as it weights equally all categories considered in the classification.

4.1. Emotion detection

In regard to emotion detection, in Table 2 we report F1 scores for all 3 models. The proposed CNN-BiLSTM-PM model achieves superior results compared with the other two approaches. It outperforms the BERT model by 4 percentage points (6%), and the vector-id model by 14 percentage points (26%). Furthermore, the proposed CNN-BiLSTM model is designed for multi-label prediction instead of single emotion detection as in the case of the BERT model.

Model	F1-macro (%)
CNN-BiLSTM-PM	**67.68 (0.55)**
CNN-BiLSTM-vector-id	53.72 (0.26)
BERT-GoEmotions-majority	64.00 (n/a)

Table 2. Results of the emotion recognition task (standard deviations reported in parentheses).

4.2. Abusive language detection

In regard to the cross condition, 20% of unique texts were sampled for a test set and annotations from all raters for those texts were retrieved from the full dataset. Therefore, in the evaluation phase the model is fed with text samples not seen before and has to predict labels for all raters that annotated a given text.

Results reported in Table 3 indicate that the proposed personalised model with PM features outperforms the traditional approach by almost 7 percentage points in terms of F1 scores. However, both models achieve superior results with respect to the model in the third (cross) condition.

[1]https://colab.research.google.com/drive/1x2FDbRPx9d_B9YlhP6nr1P8TJAzZp0SW?usp=sharing

Model	F1-macro (%)
CNN-BiLSTM-PM	**47.60 (2.76)**
CNN-BiLSTM-majority	40.90 (3.84)
CNN-BiLSTM-cross	38.64 (1.87)

Table 3. Performance results for abusive language detection (standard deviations reported in parentheses).

5. Discussion

We provide further evidence that the traditional "gold standard" approach to the study of highly subjective phenomena such as emotion and abusive language detection is no longer suitable in computational linguistics. We show the value of implementing a personalised approach to supervised machine learning models, in particular in highly subjective tasks. We introduce the Personalisation Metric, which significantly improves the quality of prediction of deep learning models on the one hand, and is easy to implement on the other hand. Although other features related to the users or annotators proved useful in the previous studies, for example demographic factors (Hovy, 2015), they are often difficult to obtain due to privacy issues or missing data on social media platforms, among others.

In the current study, we propose a method to model individual factors that influence the perception of highly subjective phenomena such as emotions and abusive language. The introduced Personalisation Metric is suitable for categorical labels and classification tasks (i.e., categorical labels are taken to calculate the metric score), and therefore extends previous solutions proposed for numerical labels (i.e., numerical labels are taken to calculate the metric) and regression models (see Miłkowski et al., 2021).

A personalised approach to text classification constitutes an alternative to a popular majority-based method where a machine learning model is both trained and evaluated on majority-aggregated labels obtained from a set of annotators. However, when applied to predict individual users' preferences in reality, those models perform rather poorly (Gordon et al., 2021). Thus, a new approach that takes into account not only the opinion of the majority but also individual users is needed. Others provide solutions with the use of demographic information (Hovy, 2015), we propose the one based on annotation behaviour of an individual. We demonstrate that the proposed method outperforms standard majority-based classifiers applied to both majority-aggregated and individual labels (BERT-GoEmotions-majority and CNN-BiLSTM-majority, and CNN-BiLSTM-cross models, respectively) as well as models that incorporate information about id number of annotators (CNN-BiLSTM-vector-id model). Results obtained in the current study corroborate previous findings in this area (Kocoń et al., 2021; Miłkowski et al., 2021).

Scalability of the proposed method to new users could be addressed in two ways. First, new users could be provided with a sample of data to annotate in order to compute the PM metric and define their deviation from the majority opinion. Second solution makes use of a collaborative filtering technique, broadly used in recommendation systems. Here, the PM metric could be computed as the average value from several users similar in some aspects to a new user. This similarity could regard user behaviour on a platform or demographic information such as gender or age. Future studies could examine the suitability of those two methods for the proposed personalised approach to text classification.

Although the traditional approach to text classification works well for the majority of people, the alternative acknowledges those individuals that do not have enough representation in the majority perspective to work satisfactorily for them. It concerns in particular subjective phenomena such as emotion and abusive language perception. Further research in natural language processing could examine the impact of incorporation of user-based information on classification performance and advance the field not only in terms of new state-of-the-art performance but also practical suitability for the end users (Gordon et al., 2021).

Acknowledgments

The work reported in this paper was partially supported by the Polish National Science Centre under grant 2020/39/D/HS1/00488.

References

Akhtar, S., Basile, V. and Patti, V. (2020, October). Modeling annotator perspective and polarized opinions to improve hate speech detection. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing (Vol. 8, pp. 151-154)*.

Akhtar, S., Basile, V. and Patti, V. (2021). Whose opinions matter? perspective-aware models to identify opinions of hate speech victims in abusive language detection. *arXiv preprint arXiv*:2106.15896.

Basile, V. (2020). It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *2020 AIxIA Discussion Papers Workshop* (Vol. 2776, pp. 31-40). CEUR-WS.

Binns, R., Veale, M., Kleek, M.V. and Shadbolt, N. (2017, September). Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *International conference on social informatics (pp. 405-415)*. Springer, Cham.

Cohen, J. (1960). Kappa: Coefficient of concordance. *Educ Psych Measurement*, 20(37), 37-46.

Curry, A.C., Abercrombie, G. and Rieser, V. (2021). ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. *arXiv preprint arXiv*:2109.09483.

Davani, A.M., Díaz, M. and Prabhakaran, V. (2022). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. In *Transactions of the Association for Computational Linguistics*, 10, 92-110.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G. and Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. *arXiv preprint arXiv*:2005.00547.

Eugenio, B. D., & Glass, M. (2004). The kappa statistic: A second look. *Computational linguistics*, *30*(1), 95-101.

Gordon, M. L., Zhou, K., Patel, K., Hashimoto, T., & Bernstein, M. S. (2021, May). The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-14)*.

Hovy, D. (2015, July). Demographic factors improve classification performance. In *Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing (pp. 752-762)*.

Kocoń, J., Figas, A., Gruza, M., Puchalska, D., Kajdanowicz, T. and Kazienko, P. (2021). Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. In *Information Processing & Management*, 58(5), 102643.

Larimore, S., Kennedy, I., Haskett, B. and Arseniev-Koehler, A. (2021). Reconsidering annotator disagreement about racist language: noise or signal?. *SocialNLP 2021*, 81.

Lukin, S.M., Anand, P., Walker, M. and Whittaker, S. (2017). Argument strength is in the eye of the beholder: Audience effects in persuasion. *arXiv preprint arXiv*:1708.09085.

Miłkowski, P., Gruza, M., Kanclerz, K., Kazienko, P., Grimling, D., and Kocoń, J. (2021, August). Personal bias in prediction of emotions elicited by textual opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop (pp. 248-259)*.

Pennington, J., Socher, R. and Manning, C.D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543)*.

Prabhakaran, V., Davani, A.M. and Diaz, M. (2021). On Releasing Annotator-Level Labels and Information in Datasets. In *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop* (pp. 133-138).

Rizos, G. and Schuller, B.W. (2020, June). Average jane, where art thou?–recent avenues in efficient machine learning under subjectivity uncertainty. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (pp. 42-55)*. Springer, Cham.

Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y. and Smith, N.A. (2021). Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv*:2111.07997.

Tam, S., Said, R.B. and Tanriöver, Ö.Ö. (2021). A ConvBiLSTM deep learning model-based approach for Twitter sentiment classification. In *IEEE Access*, 9, 41283-41293.

Tanana, M.J., Soma, C.S., Kuo, P.B., Bertagnolli, N.M., Dembe, A., Pace, B.T., Srikumar, V., Atkins, D.C. and Imel, Z.E. (2021). How do you feel? Using natural language processing to automatically rate emotion in psychotherapy. *Behavior Research Methods*, 53(5), 2069-2082

Wang, X., Jiang, W. and Luo, Z. (2016, December). Combination of convolutional and recurrent neural network for sentiment analysis of short texts. In *Proceedings of COLING 2016* (pp. 2428-2437).

Waseem, Z. (2016, November). Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. In *Proceedings of the first workshop on NLP and computational social science* (pp. 138-142).

Wich, M., Bauer, J. and Groh, G. (2020, November). Impact of politically biased data on hate speech classification. In *Proceedings of the Fourth Workshop on Online Abuse and Harms* (pp. 54-64).