Recognition of emotions, valence and arousal in large-scale multi-domain text reviews

Jan Kocoń*, **Arkadiusz Janz***, **Piotr Miłkowski***, **Monika Riegel**†, **Małgorzata Wierzba**†,
Artur Marchewka†, Agnieszka Czoska‡, Damian Grimling ‡,
Barbara Konat‡¡, Konrad Juszczyk‡¡, Katarzyna Klessa¡, Maciej Piasecki*

*Wroclaw University of Science and Technology
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
{jan.kocon, arkadiusz.janz, piotr.milkowski, maciej.piasecki}@pwr.edu.pl

† Laboratory of Brain Imaging, Nencki Institute of Experimental Biology of Polish Academy of Sciences
Ludwika Pasteura 3, 02-093 Warszawa
{a.marchewka, m.riegel, m.wierzba}@nencki.gov.pl

‡W3A.PL Sp. z o.o.
Piątkowska 110A/1, 60-649 Poznań, Polska
{agnieszka, damian, barbara, konrad}@sentimenti.pl

¡Adam Mickiewicz University, Faculty of Modern Languages and Literatures
Niepodległości 4, 61-874 Poznań
{klessa}@amu.edu.pl

Abstract

In this article, we present a novel multidomain dataset of Polish text reviews. The data were annotated as part of a large study involving over 20,000 participants. A total of 7,000 texts were described with metadata, each text received about 25 annotations concerning polarity, arousal and eight basic emotions, marked on a multilevel scale. We present a preliminary approach to data labelling based on the distribution of manual annotations and to the classification of labelled data using logistic regression and bi-directional long short-term memory recurrent neural networks.

1. Introduction

Emotions are a crucial part of natural human communication, conveyed by both what we say and how we say it. In this study, we focus on emotions attributed by Polish native speakers to written Polish texts. The results presented in this paper combine machine learning with an empirical approach to language and emotions expressed verbally.

Introduction of machine learning (ML) to the area of text mining resulted in the rapid growth of the field in recent years. However, an automatic emotion recognition with Machine Learning remains a challenging task due to the scarcity of high quality and large scale data sources. Numerous approaches were attempted to annotate words concerning their polarity and emotions for various languages (Riegel et al., 2015). Such datasets, however, are limited in size, typically consisting of several thousands of words, while lexicons are known to be much bigger.[1] The size of the known and annotated affective word lists constrains their usage in natural language processing.

In emotion research, words are usually characterised according to two dominant theoretical approaches to the nature of emotion: dimensional account and categorical account. According to the first account proposed in (Russell and Mehrabian, 1977), each emotion state can be represented by its location in a multidimensional space, with valence or polarity (negativity/positivity) and arousal (low/high) explaining most of the observed variance. In the competing account, several basic or elementary emotion states are distinguished, with more complex, subtle emotion states emerging as their combination. To categorise emotions, semantic concepts drawn from natural language are used, as corresponding to particular behavioural or physiological response patterns. The concept of basic emotions itself has been interpreted in various ways, and thus different theories posit different numbers of categories of emotion, with (Ekman, 1992) and (Plutchik, 1982) gaining most recognition in the scientific community.

On the other hand, the most popular approach in natural language processing, but also in applied usages of emotion annotation is sentiment analysis which takes into account only polarity (negativity/positivity). It is understandable since the emotion annotation of textual data faces difficulties in the two conventional approaches to annotation. In the first approach, a small number (usually 2 to 5) trained annotators are engaged and because of the differences between individual opinions, enhanced by multiple choice possibilities (most commonly 6 or 8 emotions), may lead to poor results of inter-annotator agreement (Hripcsak and Rothschild, 2005). The other approach, based on crowd annotations on platforms such as Amazon Turk (Paolacci and Chandler, 2014) leads to a similar problem: the inter-labeller variability of annotations is high because such plat-

[1] The largest dictionary of English, Oxford English Dictionary, for example, contains around 600,000 words in its online version `https://public.oed.com/about`

forms are open to users of different nationalities while only the native speakers of a given language can distinguish subtleties of emotional connotations.

In this study, we applied an approach that proved useful in previous experiments (Riegel et al., 2015). Thus, our annotation schema follows the account of Russel and Mehrabian, as well as those proposed by Ekman or Plutchik. Finally, by combining simple annotation schema with crowd annotation, we were able to effectively acquire a large amount of data, while at the same time preserving the high quality of the data. Sentiment analysis enhanced with eight basic emotions leads to new possibilities of studying people's attitudes towards brands, products and their features, political views, movie or music choices or financial decisions, including stock exchange activity. Moreover, comparing the results of meaning and text ranking leads to a better understanding of text processing, especially constructing the emotional meaning of texts by readers.

2. Data annotation

To create Sentimenti database, a total of over 20,000 unique respondents (with approximately equal number of male and female participants) was sampled from Polish population (sex, age, native language, place of residence, education level, marital status, employment status, political beliefs and income were controlled, among other factors). To collect the data, a combined approach of different methodologies was used, namely: Computer Assisted Personal Interview (CAPI) and Computer Assisted Web Interview (CAWI).

The annotation schema was based on procedures most widely used in previous studies aiming to create the first datasets of Polish words annotated in terms of emotion (NAWL, (Riegel et al., 2015); NAWL BE, (Wierzba et al., 2015); plWordNet-emo (Zaśko-Zielińska et al., 2015; Janz et al., 2017)). Thus, we collected extensive annotations of valence (polarity), arousal, as well as eight emotion categories: joy, sadness, trust, disgust, fear, anger, surprise and anticipation.

The total number of over 30,000 word meanings from Polish WordNet (Piasecki et al., 2009) was annotated, with each meaning ranked at least 50 times on each scale. The selection of word meanings was based on the results of the plWordNet-emo (Zaśko-Zielińska et al., 2015) project, in which linguists annotated over 87K lexical units with over 178K annotations containing information about emotions, valence (polarity) and valuations (statistics from May 2019). At the time when the selection was made (July 2017) 84K annotations were covering 54K word meanings and 41K synsets. We observed that 27% of all annotations (23K) were not neutral. The number of synsets having lexical units with polarity different than neutral was 9K. We have adopted the following assumptions for the selection procedure:

- word meanings that we know are not neutral are more important,
- polarity sign of the synset is the polarity sign of word meanings within the synset (valid in 96% of cases),
- the maximum number of selected word meanings from the same synset is 3,
- the degree of synsets (treated as nodes in plWordNet graph) which are sources of selected word meanings should be in range $[3, 6]$.

Word meanings were presented to respondents as collocations manually prepared by linguists.

Moreover, in a follow-up study, a total number of over 7,000 texts (short phrases or paragraphs of text) were annotated in the same way, with each text assessed at least 25 times on each scale. Before attempting the assessment task, subjects were instructed to rank word meanings rather than words, as well as encouraged to indicate their immediate, spontaneous reactions. Participants had unlimited time to complete the task, and they were able to quit the assessment session at any time and resume their work later on. The final collection of texts for emotive annotation was acquired from Web reviews of two distinct domains: *medicine*[2] (2000 reviews) and *hotels*[3] (2000 reviews). Due to the scarcity of neutral reviews in these data sources, we decided to acquire yet another sample from potentially neutral Web sources being thematically consistent with selected domains, i.e. medical information sources [4] (500 paragraphs) and hotel industry news [5] (500 paragraphs). The phrases for annotation were extracted using *lexico-semantic-syntactic patterns* (LSS) manually created by linguists to capture one of the four effects affecting sentiment: *increase, decrease, transition, drift*. Most of these phrases belong to previously mentioned thematic domains. The source for the remaining phrases were Polish WordNet glosses and usage examples (Piasecki et al., 2009).

3. Data transformation

We decided to carry out the recognition of specific dimensions as a classification task. Eight basic emotions were annotated by respondents on a scale of integers from range $[0, 4]$ and the same scale was also used for arousal dimension. For valence dimension, a scale of integers from range $[-3, 3]$ was proposed to obtain a more clear gradation of effect size. We divided the valence scores into two groups: positive (valence_p) and negative (valence_n). This division results from the fact that there were texts that received scores from both polarities. We wanted to keep that distribution (see Algorithm 1). For the rest of dimensions, we assigned the average value of all scores (normalised to the range $[0, 1]$) to the text.

3.1. Scores distribution

As a part of this study, a collection of 7004 texts was annotated. To investigate the underlying empirical distribution of emotive scores we analysed our data concerning each dimension separately. We performed two statistical tests to verify the multimodality of scores distribution in our sample for each dimension. The main purpose of this

[2] www.znanylekarz.pl

[3] pl.tripadvisor.com

[4] naukawpolsce.pap.pl/zdrowie

[5] hotelarstwo.net, www.e-hotelarstwo.com

Algorithm 1 Estimating the average value of positive and negative valence for a single review.

```
Require: $V$: list of all valence scores;
    $m = 3$: the maximum absolute value of polarity;
Ensure: Pair $(p, n)$ where $p$ is average positive valence,
    and $n$ is average negative valence;
$(p, n) = (0, 0)$
for $v \in V$ do
    if $v < 0$ then $n = n + |v|$ else $p = p + v$
return $\big(p \div (|V| \cdot m), n \div (|V| \cdot m)\big)$
```

analysis was to identify if there exists a specific decision boundary splitting our data into distinct clusters, to separate the examples sharing the same property (e.g. positive texts) from the examples that do not share this property (e.g. non-positive texts). The first test was Hartigans' dip test. It uses the maximum difference for all averaged scores, between the empirical distribution function, and also the unimodal distribution function that minimises the maximum difference (Hartigan et al., 1985). There are the unimodal null hypothesis and a multimodal alternative. The second one is Silverman's mode estimation test which uses kernel density estimation methods to examine the number of modes in a sample (Silverman, 1981). If the null hypothesis of unimodality ($k = 1$) was rejected, we also tested if there are two modes ($k = 2$) or more (Neville and Brownstein, 2018). We used *locmodes* R package to apply statistical testing (Ameijeiras-Alonso et al., 2016) with Hartigans' and Silverman's tests on our annotation data. For all dimensions we could not reject the null hypothesis of bimodality and only in 2 cases (arousal, disgust) we could reject the null hypothesis of unimodality by the result of both tests (see Table 1).

Dimension	SI_mod1	SI_mod2	HH
valence_n	0.000	0.812	0.000
valence_p	0.000	0.460	0.000
arousal	0.340	0.606	0.118
joy	0.000	0.842	0.000
sadness	0.000	0.424	0.000
fear	0.892	0.674	0.032
disgust	0.784	0.500	0.178
surprise	0.288	0.360	0.000
anticipation	0.522	0.321	0.000
trust	0.034	0.736	0.226
anger	0.000	0.630	0.000

Table 1: p–values for Silverman's test with $k = 1$ (SI_mod1), $k = 2$ (SI_mod2) and Hartigans' dip test (HH).

The distributions of averaged scores for all texts are presented in Figure 1. We decided to partition all scores for each dimension into two clusters using k-means clustering (Hartigan and Wong, 1979). Clusters are represented in Figure 1 with different colours. We assign a label (corresponding to the dimension) if the score for the dimension is higher than the threshold determined by k-means. Each review may be described with multiple labels.

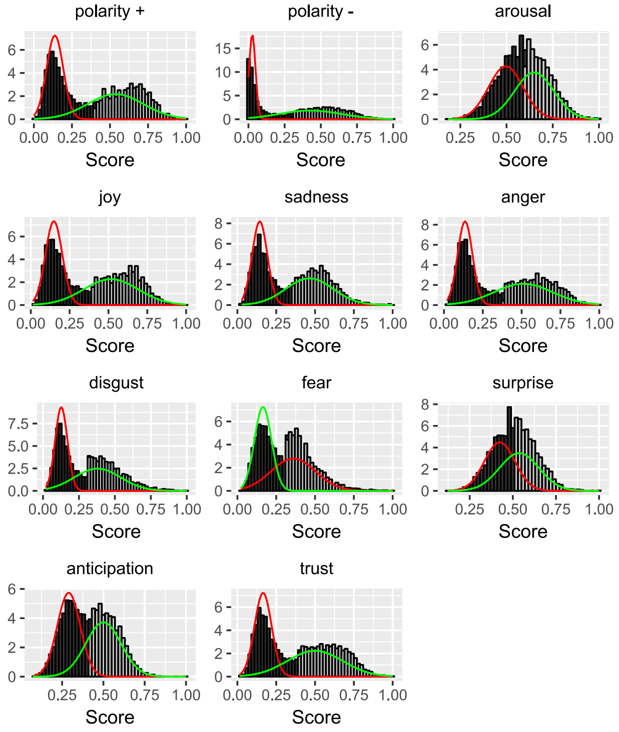

Figure 1: Distribution of avg. scores for all dimensions.

4. Experiments

In our experimental part, we decided to use a popular baseline model based on fastText algorithm (Bojanowski et al., 2017; Joulin et al., 2017) as a reference method for the evaluation. FastText's supervised models were used in many NLP tasks, especially in the area of sentiment analysis, e.g. for hate speech detection (Badjatiya et al., 2017), emotion and sarcasm recognition (Felbo et al., 2017) or aspect-based sentiment analysis in social media (Wojatzki et al., 2017). The unsupervised fastText models were also used to prepare word embeddings of Polish (see Section 4.1.). In our experiments, we used supervised fastText models as a simple multi-label text classifier for sentiment and emotion recognition. We used *one-versus-all cross-entropy* loss and 250 training epochs, with KGR10 pretrained word vectors (Kocoń and Gawor, 2019) (described in Section 4.1.) for all evaluation cases.

In recent years deep neural networks have begun to dominate natural language processing (NLP) field. The most popular solutions incorporate bidirectional long short-term memory neural networks (henceforth BiLSTM). BiLSTM-based approaches were mainly applied in the information extraction area, e.g. in the task of proper names recognition, where the models are often combined with conditional random fields (CRF) to impose additional constraints on sequences of tags as presented in (Habibi et al., 2017).

LSTM networks have proved to be very effective in sentiment analysis, especially for the task of polarity detection (Wang et al., 2016; Baziotis et al., 2017; Ma et al., 2018). In this study, we decided to adopt the multi-labelled BiLSTM networks and expand our research to the more challenging task of emotion detection. As an input for BiLSTM networks we used pre-trained fastText embeddings trained on KGR10 corpus (Kocoń and Gawor, 2019). The parameters used for training procedure were as follows: `MAX_WORDS=128` (94% of reviews have 128 words or less), `HIDDEN_UNITS=1024`, `DROPOUT_RATIO=0.2`, `EPOCHS=250`, `OPTIMIZER=ADAM`, `LEARNING_RATE=0.001`, `BATCH_SIZE=128`.

4.1. Word embeddings

The most popular text representations in recent machine learning solutions are based on *word embeddings*. Dense vector space representations follow the distributional hypothesis that the words with similar meaning tend to appear in similar contexts. Word embeddings capture the similarity between words and are often used as an input for the first layer of deep learning models. *Continuous Bag-of-Words* (CBOW) and *Skip-gram* (SG) models are the most common methods proposed to generate distributed representations of words embedded in a continuous vector space (Mikolov et al., 2013).

With the progress of machine learning methods, it is possible to train such models on larger data sets, and these models often outperform the simple ones. It is possible to use a set of text documents containing even billions of wOur article is based on this work in the development of experiments and we are researching texts from similar domains.ords as training data. Both architectures (CBOW and SG) describe how the neural network learns the vector representations for each word. In CBOW architecture the task is *predicting the word given its context*, and in SG the task is *predicting the context given the word*.

Numerous methods have been developed to prepare vector space representations of words, phrases, sentences or even full texts. The quality of vector space models depends on the quality and the size of the training corpora used to prepare the embeddings. Hence, there is a strong need for proper evaluation metrics, both intrinsic and extrinsic (task-based evaluation), to evaluate the quality of vector space representations including word embeddings (Schnabel et al., 2015), (Piasecki et al., 2018). Pre-trained word embeddings built on various corpora are already available for many languages, including the most representative group of models built for English (Kutuzov et al., 2017) language.

In (Kocoń and Gawor, 2019) we introduced multiple variants of word embeddings for Polish built on KGR10 corpora. We used the implementation of CBOW and Skip-gram methods provided with fastText tool (Bojanowski et al., 2017). These models are available under an open license in the CLARIN-PL project repository[6]. With these embeddings, we obtained a favourable results in two NLP tasks: recognition of temporal expressions (Kocoń and Gawor, 2019) and recognition of named entities (Marcińczuk et al., 2018). For this reason, the same model of word embeddings was used for this work, which is *EC1* (Kocoń and Gawor, 2019) (`kgr10.plain.skipgram.dim300.neg10.bin`).

4.2. Evaluation procedure

We prepared three evaluation scenarios to test the performance of fastText and BiLSTM baseline models. The most straightforward scenario is a single domain setting (SD) where the classifier is trained and tested on the data representing the same thematic domain. In a more realistic scenario, the thematic domain of training data differs from the application domain. This means that there may exist a discrepancy between feature spaces of training and testing data which leads to a significant decrease of classifier's performance in the application domain. To test the classifier's ability to bridge the gap between source and target domains we propose a second evaluation scenario called 1-Domain-Out (DO). This scenario is closely related to the task of unsupervised domain adaptation (UDA), where we focus on transferring the knowledge from labelled training data to unlabelled testing data. The last evaluation scenario is a multidomain setting where we merge all available labelled data representing different thematic domains into a single training dataset (MD).

- *Single Domain, SD* – train/dev/test sets are from the same domain (3 settings, metric: F1-score).
- *1-Domain-Out, DO* – train/dev sets are from two domains, test set is from the third domain (3 settings, metric: F1-score).
- *Mixed Domains, MD* – train/dev/test sets are randomly selected from all domains (1 setting, metrics: precision, recall, F1-score, AUC_ROC).

We prepared seven evaluation settings with a different domain-based split of the initial set of texts. The final division is presented in Table 2.

Type	Setting	Train	Dev	Test	SUM
SD	Hotels	2504	313	313	3130
	Medicine	2352	293	293	2938
	Other	750	93	93	936
DO	Hotels-Other	3660	406	-	4066
	Hotels-Medicine	5462	606	-	6068
	Medicine-Other	3487	387	-	3874
MD	All	5604	700	700	7004

Table 2: The number of texts in the evaluation settings.

To tune our baseline methods we decided to use a *dev* set. We calculated the optimal decision threshold for each dimension using receiver operating characteristic (ROC) curve, taking the threshold which produces the point on ROC closest to $(FPR, TPR) = (0, 1)$.

5. Results

Table 3 shows the results for *SD* evaluation. There are 11 results for each of the 3 domains. BiLSTM classifier outperformed FastText in 27 out of 33 cases. Table 4 shows the results for *DO* evaluation. Here BiLSTM classifier provided better quality for 31 out of 33 cases. The last *MD* evaluation results are in Table 5 (P, R, F1-score) and Figure 2 (ROC). BiLSTM outperformed FastText in 31 out of

[6] `https://clarin-pl.eu/dspace/handle/11321/606`

36 cases (Table 5). ROC_AUC is the same for both classifiers in 4 cases (2 of them are micro and macro-average ROC). For the rest of the curves, BiLSTM outperformed FastText in 7 out of 9 cases. The most interesting phenomenon can be observed in Table 4 where the differences are the greatest. This may indicate that the deep neural network was able to capture domain-independent features (pivots) which is an important ability for domain adaption tasks.

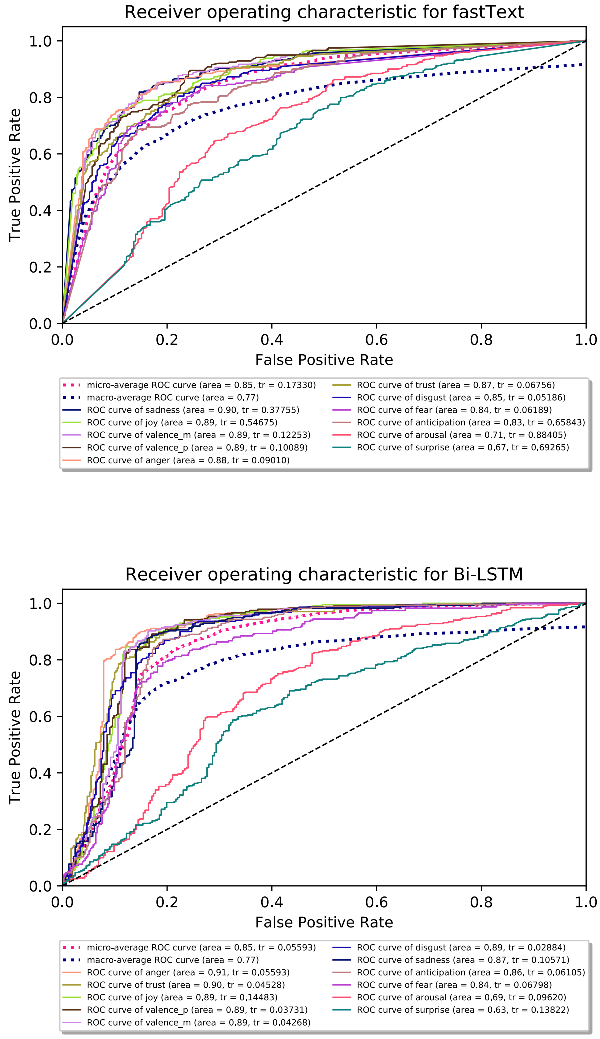

Figure 2: ROC curves for FastText and BiLSTM classifiers.

6. Conclusions

In this preliminary study, we focused on basic neural language models to prepare and evaluate baseline approaches to recognise emotions, valence and arousal in multi-domain textual reviews. Further plans include the evaluation of hybrid approaches combining machine learning approaches and lexico-syntactic rules augmented with semantic analysis of word meanings. We also plan to automatically expand the annotations of word meanings to the rest of lexical units within plWordNet using the propagation methods presented in (Kocoń et al., 2018a; Kocoń et al., 2018b). We intend to test other promising methods later, such as Google BERT (Devlin et al., 2018), OpenAI GPT-2 (Radford et al., 2019) and domain dictionaries construction methods utilising WordNet (Kocoń and Marcińczuk, 2016).

Automatic emotion annotation has both scientific and applied value. Modern business is interested in the opinions, emotions and values associated with brands and products. Retailers and merchants collect vast amounts of customer feedback and rumours both from in-store and posted online. What is more, relation departments monitor the impact of their campaigns and need to know whether it was positive and touching for customers. In this context, the results of monitoring opinions, reactions, and emotions present great value, because they fuel decisions and behaviour (Tversky and Kahneman, 1989). However, most of the existing solutions are still limited to manual annotation and simplified methods of analysis.

The large database built in the Sentimenti project covers a wide range of Polish vocabulary and introduces an extensive emotive annotation of word meanings in terms of their polarity, basic emotions and affective arousal. The results of such research can be used in several applications – media monitoring, chatbots, stock prices forecasting, search engine optimisation for advertisements and other types of content. In the last decades, the development of Internet services gave us an unprecedented amount of data, resulting in the *big data revolution* (Kitchin, 2014). This also includes the textual data coming directly from social media and other sources.

We also provide a preliminary overview of ML methods for automatic analysis of people's opinions in terms of expressed emotions and their attitudes. Since the participants of our CAPI and CAWI studies represent a wide cross-section of population we can adapt our methods to specific target groups of people. This introduces the much needed human aspect to artificial intelligence and machine learning in natural language processing.

7. Data availability

Due to the commercial nature of the Sentimenti project, it is planned to make 10% of the project data available soon. The data will be published at www.sentimenti.pl. We will consider making more data accessible in the future.

Acknowledgements

Co-financed by the Polish Ministry of Education and Science, CLARIN-PL Project and by the National Centre for Research and Development, Poland, grant no POIR.01.01.01-00-0472/16 – *Sentimenti* (`http://w3a.pl/projekty/`).

8. References

Ameijeiras-Alonso, Jose, Rosa M Crujeiras, and Alberto Rodríguez-Casal, 2016. Mode testing, critical bandwidth and excess mass. *TEST*:1–20.

Setting	*Classifier*	*Valence*$_p$	*Valence*$_n$	*Arousal*	*Joy*	*Surprise*	*Anticipation*	*Trust*	*Sadness*	*Anger*	*Fear*	*Disgust*
1. Hotels	FastText	90.53	88.43	66.67	89.08	62.63	77.91	83.41	86.04	88.33	65.81	81.86
	BiLSTM	89.74	89.54	67.66	86.84	46.62	82.11	80.83	88.46	89.54	63.53	82.76
2. Medicine	FastText	75.37	56.18	61.54	75.00	62.00	75.49	74.14	64.32	59.09	45.90	73.20
	BiLSTM	82.18	82.40	65.31	84.15	64.38	80.31	82.47	86.33	85.23	83.04	74.04
3. Other	FastText	66.67	66.67	62.34	62.86	48.57	51.52	45.28	77.27	48.15	45.28	46.51
	BiLSTM	80.52	75.95	65.17	80.49	33.90	64.71	70.37	79.52	65.52	68.66	62.75

Table 3: F1-scores for *Single Domain* evaluation. (Train, Dev, Test) sets for settings are the same as in Table 2, rows 1-3.

Setting	*Classifier*	*Valence*$_p$	*Valence*$_n$	*Arousal*	*Joy*	*Surprise*	*Anticipation*	*Trust*	*Sadness*	*Anger*	*Fear*	*Disgust*
4. Hotels-Other vs Medicine	FastText	61.44	72.79	63.08	61.73	59.03	58.10	65.54	75.27	71.97	71.33	63.20
	BiLSTM	74.56	76.61	66.00	71.25	62.62	70.32	67.52	80.40	73.97	74.03	69.80
5. Hotels-Medicine vs Other	FastText	61.05	39.29	37.50	65.96	20.51	45.95	42.42	25.45	05.71	17.65	48.65
	BiLSTM	73.17	56.34	35.29	75.00	51.52	60.53	56.67	61.90	43.48	57.69	48.39
6. Medicine-Other vs Hotels	FastText	73.93	78.26	35.18	71.86	56.32	73.25	73.45	72.96	76.60	50.96	71.21
	BiLSTM	88.89	87.07	51.88	87.07	62.07	84.76	82.79	86.14	87.14	63.44	82.57

Table 4: F1-scores for *1-Domain-Out* evaluation. (Train/Dev, Test) sets (see Table 2) for these settings are: 4. (Hotels-Other.Train/Dev, Medicine.Test), 5. (Hotels-Medicine.Train/Dev, Other.Test), 6. (Medicine-Other.Train/Dev, Hotels.Test).

Dim.	FastText			BiLSTM		
	P	R	F	P	R	F
Valence$_p$	73.41	77.41	75.36	77.61	84.10	80.72
Valence$_n$	75.79	87.00	81.01	81.31	89.53	85.22
Arousal	67.48	69.16	68.31	67.09	66.04	66.56
Joy	70.61	81.14	75.51	77.51	84.65	80.92
Surprise	65.07	64.31	64.69	67.67	59.88	63.54
Anticip.	72.28	77.66	74.78	79.66	81.91	80.77
Trust	65.32	79.02	71.52	73.91	82.93	78.16
Sadness	81.73	82.55	82.14	83.88	85.57	84.72
Anger	80.92	78.52	79.70	82.03	89.63	85.66
Fear	69.20	77.78	73.24	68.84	81.20	74.51
Disgust	66.80	77.73	71.85	71.71	84.09	77.41
Avg.	71.69	77.48	74.38	75.57	80.87	78.02

Table 5: Precision, recall and F1-score for *Mixed Domains* evaluation.

Badjatiya, Pinkesh, Shashank Gupta, Manish Gupta, and Vasudeva Varma, 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee.

Baziotis, Christos, Nikos Pelekis, and Christos Doulkeridis, 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*.

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

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ekman, Paul, 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200.

Felbo, Bjarke, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann, 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*.

Habibi, Maryam, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser, 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48.

Hartigan, John A, Pamela M Hartigan, et al., 1985. The dip test of unimodality. *The annals of Statistics*, 13(1):70–84.

Hartigan, John A and Manchek A Wong, 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Hripcsak, George and Adam S. Rothschild, 2005. Technical Brief: Agreement, the F-Measure, and Reliability in Information Retrieval. *JAMIA*, 12(3):296–298.

Janz, Arkadiusz, Jan Kocoń, Maciej Piasecki, and Zaśko-Zielińska Monika, 2017. plWordNet as a Basis for

Large Emotive Lexicons of Polish. In *LTC'17 8th Language and Technology Conference*. Poznań, Poland: Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov, 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics.

Kitchin, Rob, 2014. *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

Kocoń, Jan, Arkadiusz Janz, and Maciej Piasecki, 2018a. Classifier-based Polarity Propagation in a Wordnet. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18)*.

Kocoń, Jan, Arkadiusz Janz, and Maciej Piasecki, 2018b. Context-sensitive Sentiment Propagation in WordNet. In *Proceedings of the 9th International Global Wordnet Conference (GWC'18)*.

Kocoń, Jan and Michal Gawor, 2019. Evaluating KGR10 Polish word embeddings in the recognition of temporal expressions using BiLSTM-CRF. *CoRR*, abs/1904.04055.

Kocoń, Jan and Michał Marcińczuk, 2016. Generating of Events Dictionaries from Polish WordNet for the Recognition of Events in Polish Documents. In *Text, Speech and Dialogue, Proceedings of the 19th International Conference TSD 2016*, volume 9924 of *Lecture Notes in Artificial Intelligence*. Brno, Czech Republic: Springer.

Kutuzov, Andrei, Murhaf Fares, Stephan Oepen, and Erik Velldal, 2017. Word vectors, reuse, and replicability: Towards a community repository of large-text resources. In *Proceedings of the 58th Conference on Simulation and Modelling*. Linköping University Electronic Press.

Ma, Yukun, Haiyun Peng, and Erik Cambria, 2018. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Marcińczuk, Michał, Jan Kocoń, and Michał Gawor, 2018. Recognition of Named Entities for Polish-Comparison of Deep Learning and Conditional Random Fields Approaches. In *Proceedings of PolEval 2018 Workshop*. Warsaw, Poland: Institute of Computer Science, Polish Academy of Sciences.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*.

Neville, Zachariah and Naomi C Brownstein, 2018. Macros to conduct tests of multimodality in SAS. *Journal of Statistical Computation and Simulation*, 88(17):3269–3290.

Paolacci, Gabriele and Jesse Chandler, 2014. Inside the turk: Understanding mechanical turk as a participant pool. *Current Directions in Psychological Science*, 23(3):184–188.

Piasecki, Maciej, Bernd Broda, and Stanislaw Szpakowicz, 2009. *A wordnet from the ground up*. Oficyna Wydawnicza Politechniki Wrocławskiej Wrocław.

Piasecki, Maciej, Gabriela Czachor, Arkadiusz Janz, Dominik Kaszewski, and Paweł Kędzia, 2018. Wordnet-based Evaluation of Large Distributional Models for Polish. In *Proceedings of the 9th Global WordNet Conference (GWC 2018)*.

Plutchik, Robert, 1982. A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5):529–553.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, 2019. Language models are unsupervised multitask learners. *OpenAI Blog*:8.

Riegel, Monika, Małgorzata Wierzba, Marek Wypych, Łukasz Żurawski, Katarzyna Jednoróg, Anna Grabowska, and Artur Marchewka, 2015. Nencki Affective Word List (NAWL): the cultural adaptation of the Berlin Affective Word List–Reloaded (BAWL-R) for Polish. *Behavior Research Methods*, 47(4):1222–1236.

Russell, James A and Albert Mehrabian, 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3):273 – 294.

Schnabel, Tobias, Igor Labutov, David M Mimno, and Thorsten Joachims, 2015. Evaluation methods for unsupervised word embeddings. In *EMNLP*.

Silverman, Bernard W, 1981. Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43(1):97–99.

Tversky, Amos and Daniel Kahneman, 1989. Rational choice and the framing of decisions. In *Multiple Criteria Decision Making and Risk Analysis Using Microcomputers*. Springer, pages 81–126.

Wang, Yequan, Minlie Huang, Li Zhao, et al., 2016. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*.

Wierzba, M., M. Riegel, M. Wypych, K. Jednorwóg, P. Turnau, A. Grabowska, and A. Marchewka, 2015. Basic emotions in the nencki affective word list (NAWL be): New method of classifying emotional stimuli. *PLoS ONE*, 10(7).

Wojatzki, Michael, Eugen Ruppert, Sarah Holschneider, Torsten Zesch, and Chris Biemann, 2017. Germeval 2017: Shared task on aspect-based sentiment in social media customer feedback. *Proceedings of the GermEval*:1–12.

Zaśko-Zielińska, Monika, Maciej Piasecki, and Stan Szpakowicz, 2015. A large wordnet-based sentiment lexicon for Polish. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*.