Multi-level analysis and recognition of the text sentiment on the example of consumer opinions

Jan Kocoń
Wrocław University
of Science and Technology
Wrocław, Poland
jan.kocon
@pwr.edu.pl

Monika Zaśko-Zielińska
University of Wrocław
Institute of Polish Studies
Wrocław, Poland
monika.zasko-zielinska
@uwr.edu.pl

Piotr Miłkowski
Wrocław University
of Science and Technology
Wrocław, Poland
piotr.milkowski
@pwr.edu.pl

Abstract

In this article, we present a novel multi-domain dataset of Polish text reviews, annotated with sentiment on different levels: sentences and the whole documents. The annotation was made by linguists in a 2+1 scheme (with inter-annotator agreement analysis). We present a preliminary approach to the classification of labelled data using logistic regression, bidirectional long short-term memory recurrent neural networks (BiLSTM) and bidirectional encoder representations from transformers (BERT).

1 Introduction

Linguistic research on sentiment recognition involves two approaches: from the perspective of analysing the occurrence of emotional words and from the perspective of the entire document. The first attempt is usually a consequence of the creation of the sentiment lexicon, e.g. manual annotation of the WordNet (Baccianella et al., 2010). The second results from the analysis of the specific text content in which we see that the sentiment of a word or phrase changes under the influence of the surrounding context (Taboada et al., 2008). This change may vary depending on the domain of the text. As a research material for our research we have chosen online customer reviews from four domains:

- *S – School* – students' reviews on the lecturers[1],
- *M – Medicine* – patients' opinions on doctors[2],
- *H – Hotels* – customer reviews of hotels[3],
- *P – Products* – buyers' opinions on products[4].

In the introduction we focus mainly on the influence of discourse on the classification of the document sentiment. We only briefly present an approach based on the analysis of emotional words. The rest of the article concerns the description of the corpus used in our analysis, guidelines for the description of the text with sentiment for annotators, the results of the pilot stage of annotation, the proper annotation and experiments with automatic recognition of the text polarity.

2 Related work

One approach for recognising polarity of text is to use a dictionary of emotional words – sentiment lexicons, e.g. WordNet annotated with polarity (Kamps et al., 2004; Takamura et al., 2005) and emotions (Janz et al., 2017). Usually the task is to determine the number of occurrence of such words with a specific polarity in the text or use a simple bag of words method (Wang and Manning, 2012). Such a solution has a number of limitations: simple methods cannot cope with irony, sarcasm, negation and more complex text structures that modify the sound of the words that make them up (Wallace et al., 2015).

The second method of analysing text polarity is examination of the sentence level with evaluating each sentence in isolation. This procedure can be supported by the external corpus of labelled sentences (Pang and Lee, 2004; Wilson et al., 2005). Nevertheless, even within one sentence we can sometimes observe several features of the analysed entity and each of them can be assessed dif-

[1] https://polwro.pl/
[2] https://www.znanylekarz.pl/
[3] https://pl.tripadvisor.com/
[4] https://www.ceneo.pl/

Proceedings of Recent Advances in Natural Language Processing, pages 559–567,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_066

ferently. It is advantageous in opinion mining to achieve both an overall opinion and specific information on the reviewed entity and its aspects. It gives us not only a general opinion on the product but allows us to notice a detailed view on the quality of the product.

This year, the first results of the Sentimenti[5] project were published, which aimed to create methods of analysing the content written on the Internet in terms of emotions expressed by the authors of the texts and the emotional impact of the readers. Within the project, a large database has been created, in which 30,000 lexical units from plWordNet database and 7,000 texts were evaluated, most of which are consumer reviews from the domain of hotels and medicine. The elements were evaluated by 20,000 unique Polish respondents in the Computer Assisted Personal Interview survey and more than 50 marks were obtained for each element, which gives more than 1.8 million annotations. Within each mark, polarisation of the element, stimulation and basic emotions aroused by the recipients are determined. The first results concerning the automatic recognition of sound and emotions for this set are presented in (Kocoń et al., 2019). Our article is based on this work in the development of experiments and we are researching texts from similar domains, but using more complex classification methods as described in Section 4. The annotation guidelines for linguists in the task of sentiment analysis on two levels were also developed: text level and sentence level (presented in Section 3).

2.1 From word level to discourse in text polarity analysis

A discourse perspective in sentiment analysis is an attempt to address limitations of previous methods (e.g. problems with negation, focusing on adjectives). It used findings of Rhetorical Structure Theory (Mann and Thompson, 1988). The attempt bears in mind local and global orientation in the text, discourse structure or topicality (Taboada et al., 2008). It allows the researcher to extract the most important sentences from the text in the perspective of the entire discourse context: nucleus satellite method (Wang et al., 2012). The relevance of the sentences is evaluated in relation to the main topic and the analysis omits some less important parts of the text. In Section 3 we present how the genre structure of a customer review affects the text sentiment polarity. It is an enhancement of the discourse perspective in sentiment analysis.

[5] https://sentimenti.com/

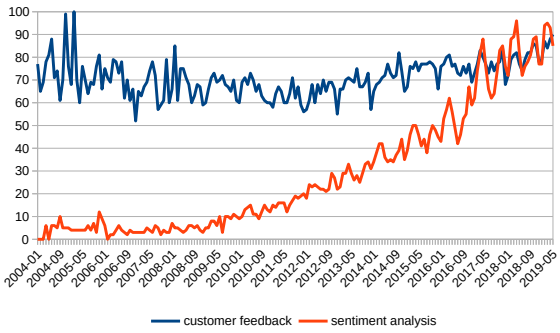

Figure 1: Google Trends (trends.google.com) data showing interest in time for search terms "customer feedback" and "sentiment analysis". On the vertical axis 100 means biggest search term popularity.

2.2 Recognition of sentiment

Sentiment analysis and opinion mining has become an interesting topic for many researches and private companies with constant growth of interest in recent years (see Figure 1), that coincides with the *big data revolution* (Kitchin, 2014). Properly evaluated data can be widely used in fields such as market analysis, public relations and product or customer feedback. Main difficulty in retrieval of those information is non-organised data, unstructured in pre-defined manner. Recently deep neural networks show relatively good performance among all available methods of processing such information (Glorot et al., 2011). Possibility of retrieving data from different sources like social networks (Pak and Paroubek, 2010), publicly available discussion boards or various marketing platforms connected with proper annotations on training data set can provide not only simple positive, negative or neutral classification but lead to accurate fine-grained sentiment prediction (Guzman and Maalej, 2014). In Section 4 we present our approach to solve this task using models based on fastText (Joulin et al., 2017), BiLSTM (Zhou et al., 2016) and BERT (Devlin et al., 2018).

3 Annotation model

Our research on sentiment analysis of customer reviews was conducted in 2018 within CLARIN-PL and it consisted of the pilot stage and the main stage. The preliminary part of analysis involved 3,000 students' opinions about lectures. It is an

authentic material provided online by students as a final assessment of the course in each subject. Each text was manually annotated by two annotators: a psychologist and a linguist, who worked accordingly to the general guidelines. In the pilot project we decided to deal with sentiment annotation of the whole text. In the sentiment annotation we used the same set of tags which is applied in *plWordNet 3.0 emo* (Zaśko-Zielińska et al., 2015; Janz et al., 2017) to lexical units: [+m] – strong positive, [+s] – weak positive, [-m] – strong negative, [-s] – weak negative, [amb] – ambiguous, [0] – neutral.

We used *amb* tag but we understood it differently. In annotation of lexical units in WordNet with sentiment *amb* indicated the possibility that units can be positive or negative in various contexts. Hence, in text sentiment analysis we assumed that *amb* denotes *ambiguous* polarity, thus the entire text cannot be clearly described by using neither positive nor negative annotation.

In the annotation we focused primarily on the strategic places in the text. In a customer review these places are the opening and closing sentences, namely the text frame. The beginning consists of the general opinion of the author on the subject of the evaluation and the end includes the author's recommendation to the recipients. The annotators created their first general evaluation based on these two segments. In the body of the review, the authors have only subtly changed these opinions. Regardless of the modification of the main opinion in the text, we did not use the *amb* tag when the text frame was unambiguously positive. The text frame polarity was influenced not only by lexical content but also nonverbal elements, e.g. emoticons or multiplication of punctuation marks.

3.1 An attempt to annotate aspects

The analysis of the content of customer reviews in our pilot project consisted of two stages: the selection of text blocks describing separate aspects and their annotation. Some parts of the text were not of an argumentative nature that could justify the author's decision to polarise the text. They included: advice (e.g. how to sign up for lectures) or general information on lectures, duration of classes, etc.

The main stage of our project was conducted based on text corpus consisting of consumer reviews (80% of texts) and texts from the corresponding domain with high probability of neutral polarity (20% of texts). We observed that the value of inter-annotator agreement in aspect annotation task was very low, below 15% of Positive Specific Agreement, PSA (Hripcsak and Rothschild, 2005).

3.2 New annotation guidelines

In the main stage of the project we decided to annotate the sentiment for the whole text (a *meta* level) and the *sentence* level. We assumed that this strategy allows to establish the acceptable value of PSA. We followed the rule that the *meta* annotation results partially from sentence annotations, however the frame polarity is the main factor for the final meta annotation. We have prepared the following rules of annotation, regardless of whether the entire text or sentence is annotated:

- SP – strong positive – entirely positive;
- WP – weak positive – generally positive, but there are some negative aspects;
- 0 – neutral;
- WN – weak negative – generally negative, but there are some positive aspects;
- SN – strong negative – entirely negative;
- AMB – ambiguous – there are both positive and negative aspects in the text that are balanced in terms of relevance.

Table 1 shows the value of Positive Specific Agreement (Hripcsak and Rothschild, 2005) obtained for a random sample of 111 documents from *Medicine* category.

4 Multi-level sentiment recognition

We selected three different classifiers for the recognition tasks:

- logistic regression (fastText) providing a baseline for text classification (Joulin et al., 2017)
- bidirectional long short-term memory recurrent network in two variants:
  - using word vector representations only
  - using the same vectors extended with general polarity information from sentiment dictionary described in Section 4.1

L	Type	Only A	A & B	Only B	PSA
M	SN	1	33	4	93%
	WN	2	2	2	50%
	0	0	24	0	100%
	AMB	1	2	3	50%
	WP	4	0	0	0%
	SP	0	31	2	97%
	sum	8	92	11	91%
S	SN	10	217	36	90%
	WN	11	1	0	15%
	0	36	273	17	91%
	AMB	2	7	14	47%
	WP	12	0	1	0%
	SP	6	194	8	97%
	sum	77	692	76	90%

Table 1: Annotation agreement between two experts (A and B) at the level (L) of text (meta – M) and sentence (S) for a sample of 111 documents using Positive Specific Agreement metric, PSA (Hripcsak and Rothschild, 2005).

- bidirectional encoder representations from transformers (BERT) with addition of sequence classification layer

We trained logistic regression model using pre-trained vectors for Polish language (Kocoń and Gawor, 2018). This approach is much faster in both training and testing than deep learning classifiers (Joulin et al., 2017), however, it has disadvantage which comes from not sharing parameters by features and classes, therefore overall result can be highly influenced by keywords with bigger class relativity.

BiLSTM on the other side takes into consideration not just words but full text fragment and basing on learnt patterns predicts potential outcome. Texts are divided into tokens and converted to corresponding word embedding vectors generated by fastText (Bojanowski et al., 2017), in this form it is possible to use it as input for neural network. Dimension of used vectors is equal to 300, therefore it must be reflected in the input shape. As a loss function for training a categorical crossentropy was chosen. Model prepared for the task consists of the following layers:

- Gaussian noise layer with standard deviation of 0.01 accepting as input shape up to 128 words with vector matrix for each word of size 300, therefore overall input shape is (128, 300)
- Bidirectional layer with LSTM instances (consisting of 1,024 hidden units using hyperbolic tangent activation method) merged with concatenation
- Dropout layer with dropout ratio equal to 0.2
- Dense layer with number of outputs representing number of all possible labels (6 in our task) using normalised exponential function (softmax) activation

BERT was designed to provide pre-trained deep bidirectional representations conditioning left and right context (Devlin et al., 2018), therefore it achieves best performance on text fragments instead of single sentences. It's architecture allows to fine-tune these representations by adding one additional output layer which suits needs of specified task. For our task as a pre-trained model `BERT-Base, Multilingual Cased`[6] was selected, which consists of 104 languages and 110M parameters, and `BertForSequenceClassification`[7] as a BERT classifier extended for multi-class classification.

4.1 Embedding vector extension

Basing on the data accommodated in plWordNet emo (Zaśko-Zielińska et al., 2015) we prepared the dictionary for all annotated lexical units and all possible levels of sentiment. Due to the lack of word sense disambiguation method, we grouped the sentiment annotations by lemmas. The final dictionary consists of a set of lemmas with assigned numbers representing the proportions of individual sentiment annotations, summing up to 1, e.g. for a lemma *akademicki* (Eng. *academic*) there were 11 annotations: 3 neutral, 4 generally negative, 3 generally positive and 1 entirely positive. Therefore arbitrary values for word "akademicki" are:

- entirely positive = 0.0909
- generally positive = 0.2727
- neutral = 0.2727
- generally negative = 0.3636
- entirely negative = 0.0000

[6] https://github.com/google-research/bert

[7] https://github.com/huggingface/pytorch-pretrained-BERT

- ambivalent = 0.0000

Using the described dictionary we have proposed additional variant of BiLSTM classifier with a word embedding vector extended with the values of sentiment for the lemma of the word from a prepared sentiment dictionary. Lemmas were retrieved during a preparation of the input data using WCRFT part-of-speech tagger (Radziszewski, 2013). Therefore, in this approach the input word vector dimension was extended with 6 values representing sentiment of the word. The final dimension of the word embedding increased from 300 to 306.

5 Evaluation

As in article (Kocoń et al., 2019), three variants of evaluation of the sentiment classification methods were prepared. The basic variant is a single domain in which the classifier is trained, tuned and tested on a set of texts from one domain. The next variant includes an analysis of the ability of the classifier to model the sentiment of the text on a level independent of the domain of the text. For this purpose, we take all available texts except the texts from the selected domain. Then the texts are divided into a training and a validation set. Testing of the model takes place on a test set from a selected domain, not taken into account at the stage of preparing the training and validation set. The third test variant allows to examine the classifiers in order to generalise the task of sentiment analysis in all available domains. For this purpose, texts from all domains are treated as one set, which is randomly divided into train, validation and test sets. Summary of the different types of evaluation:

- *SD – Single Domain* – evaluation sets created using elements from the same domain,
- *DO – Domain Out* – train/dev sets created using elements from 3 domains, test set from the remaining domain,
- *MD – Mixed Domains* – evaluation sets randomly selected from elements belonging to all domains.

Due to the fact that the data are annotated both at the level of the whole text and at the level of each sentence, a *sentence* or *text* may be an *element* in the above list. We use *SDT*, *DOT*, and *MDT* for *text* evaluation types and *SDS*, *DOS*, and *MDS* for *sentence* evaluation types. We use also prefixes of domains (*Hotels, Medicine, School, Products*) as suffixes for *SD** and *DO** variants, e.g. SDS-H is a single domain evaluation type performed on sentences within *hotels* domain, whereas DOT-M is a domain-out evaluation type performed on texts trained on texts outside *medicine* domain and tested on texts from that domain.

Table 2 shows the number of texts and sentences annotated by linguists for all evaluation types, with division into the number of elements within training, validation and test sets. Linguists annotated a total of 8,450 texts from four domains (hotels, medicine, products, school) and 35,789 sentences from two domains (hotels, medicine). The distribution of labels within each domain for texts and sentences is presented in Table 3. Average annotated text length in each domain are as follows: 788 characters in hotels, 802 in medicine, 781 in products and 442 in school.

Type	Domain	Train	Dev	Test	SUM
SDT	Hotels	2534	316	316	3166
	Medicine	2650	330	330	3310
	Products	790	98	98	986
	School	792	98	98	988
DOT	!Hotels	4756	528	-	5284
	!Medicine	4635	514	-	5149
	!Products	6727	746	-	7473
	!School	6725	746	-	7471
MDT	All	6771	846	845	8462
SDS	Hotels	12434	1554	1553	15541
	Medicine	16200	2024	2024	20248
DOS	!Hotels	16200	2024	-	18224
	!Medicine	12434	1554	-	13988
MDS	All	28581	3572	3571	35724

Table 2: The number of texts/sentences for each evaluation type in train/dev/test sets.

Type	Domain	SP	WP	0	WN	SN	AMB
SDT	Hotels	25.80	10.77	11.24	05.87	38.68	07.64
	Medicine	29.48	02.87	23.98	02.33	36.94	04.41
	Products	22.70	15.40	00.20	08.31	44.28	09.12
	School	46.92	26.19	00.20	07.99	10.11	08.59
SDS	Hotels	34.58	00.01	18.72	00.01	44.31	02.38
	Medicine	24.78	00.31	40.68	00.46	32.63	01.14

Table 3: The percentage of annotated elements in a given domain (SDT – single domain texts, SDS – single domain sentences).

6 Results

Table 4 presents the values of F1-score for each label (columns 3-8), global F1-score (column 9), micro-AUC and macro-AUC (columns 10-11) for all evaluation types related to the texts. In case of evaluation for a single domain for each label, fastText (using Logistic Regression) outperformed other classifiers in 13 out of 21 distinguishable cases. There are 12 cases for which the best score is not higher than F1=0.4. These are highly underrepresented labels, for which the part of the total annotations within the domain is less than 10% (see Table 3). The best results are obtained for *strong positive* and *strong negative* cases. Intermediate labels (*weak* and *ambiguous* variants) are much more difficult to be recognised correctly. In these cases deep neural networks outperform logisitic regression in 6 out of 11 cases. BERT classifier performs much better (13 out of 23 cases) in cross-domain knowledge transfer (DOT and MDT). For these evaluation types only 6 times fastText was better. These observations are consistent with the results of article (Kocoń et al., 2019) for *valence* dimensions.

Table 5 presents results corresponding to those presented in Table 4, but this time for sentence-level annotations. Looking at Table 3, the number of sentences marked as *weakly positive* or *weakly negative* is close to zero. These labels are not being recognised by any classifier. For other labels, regardless of the type of evaluation, the best results are mainly obtained using deep learning methods (label-specific F1-score: all 20 cases; general metrics: 12 out of 15 cases).

7 Conclusions and further steps

The automatic annotation of emotions has both a scientific and an applied value. Modern business is interested in the opinions, emotions and values associated with brands and products. Retailers and merchants collect huge amounts of customer feedback from the store and online. Moreover, the relationship departments monitor the impact of their campaigns and need to know if it was positive and affecting customers. In this context, the results of monitoring feedback, reactions and emotions are of great value as they fuel decisions and behaviors (Tversky and Kahneman, 1989). However, most of the existing solutions are still limited to manual annotations and simplified analysis methods.

Type	Classifier	SP	WP	0	WN	SN	AMB	F1	micro	macro
SDT-H	C1	**80.00**	**30.51**	93.98	00.00	**83.33**	**36.84**	**73.50**	90.87	70.20
	C2	71.11	25.00	00.00	04.76	72.44	00.00	53.00	77.14	66.13
	C3	72.82	22.95	94.12	**14.81**	81.98	27.27	68.45	89.83	72.44
	C4	71.22	10.26	**96.39**	00.00	78.16	00.00	68.45	**91.30**	**72.41**
SDT-M	C1	81.05	**15.38**	**96.39**	00.00	**80.63**	00.00	**78.55**	93.44	66.39
	C2	78.69	11.11	95.71	**14.29**	80.31	06.67	77.04	**94.02**	**71.81**
	C3	**81.93**	13.33	95.71	13.33	80.43	**07.41**	78.55	91.81	69.33
	C4	00.00	00.00	95.65	00.00	62.33	00.00	58.01	89.42	59.73
SDT-P	C1	**62.86**	27.59	00.00	**36.36**	**84.68**	16.67	**65.66**	**86.76**	**63.58**
	C2	28.57	**30.30**	00.00	00.00	69.16	**26.67**	49.49	78.37	53.46
	C3	25.00	00.00	00.00	00.00	67.26	00.00	43.43	77.64	51.46
	C4	00.00	00.00	00.00	00.00	69.74	00.00	53.54	80.40	42.60
SDT-S	C1	**79.61**	**52.63**	00.00	00.00	**50.00**	00.00	**61.62**	83.80	**62.33**
	C2	72.22	33.33	00.00	00.00	27.27	**36.36**	52.53	80.39	54.40
	C3	75.68	34.04	00.00	00.00	11.76	00.00	51.52	79.71	54.67
	C4	68.87	00.00	00.00	00.00	00.00	00.00	52.53	**83.88**	49.55
DOT-H	C1	70.91	23.08	**95.24**	00.00	**83.49**	**21.05**	**69.09**	88.21	66.34
	C2	72.73	17.02	91.76	**15.38**	78.76	16.00	65.30	88.31	71.21
	C2	73.94	19.67	88.89	10.00	75.11	16.00	62.46	87.41	70.59
	C4	**75.53**	**34.09**	90.67	00.00	82.76	00.00	68.14	**91.47**	**72.70**
DOT-M	C1	72.51	08.70	86.67	17.39	75.29	00.00	69.18	86.13	68.37
	C2	**73.17**	**22.22**	85.14	**28.57**	76.79	**16.33**	68.28	89.99	71.46
	C3	46.01	03.48	00.00	00.00	00.00	00.00	22.36	59.76	51.52
	C4	72.11	10.71	**90.32**	16.67	**85.06**	12.90	**72.51**	**90.88**	**73.14**
DOT-P	C1	60.61	47.06	00.00	**50.00**	76.00	00.00	59.60	89.71	70.29
	C2	63.16	26.09	00.00	22.22	84.91	11.76	62.63	83.19	61.73
	C3	54.55	28.57	00.00	00.00	85.71	00.00	59.18	83.81	63.25
	C4	**70.97**	**62.07**	00.00	00.00	**92.73**	**28.57**	**73.74**	**90.77**	**71.38**
DOT-S	C1	68.00	12.50	00.00	**13.33**	36.36	00.00	43.43	**90.12**	**68.67**
	C2	73.58	18.75	00.00	00.00	37.84	33.33	**51.52**	79.35	64.91
	C3	**75.25**	24.24	00.00	00.00	**41.03**	22.22	**51.52**	73.71	60.23
	C4	70.83	**29.27**	00.00	00.00	34.48	**37.50**	48.48	75.64	59.62
MDT	C1	83.20	40.27	**97.14**	10.91	85.28	17.72	76.83	88.92	68.04
	C2	81.21	41.03	96.75	09.68	83.36	21.57	74.35	92.90	74.79
	C3	81.82	00.00	96.39	**10.96**	80.75	**27.64**	72.70	87.67	74.19
	C4	**86.12**	**50.00**	94.65	00.00	**86.87**	22.86	**77.78**	**95.78**	**78.85**

Table 4: F1-scores for text-oriented evaluation. Training sets for evaluation types are the same as in Table 2 rows 1-9. Classifiers: C1 - fastText, C2 - BiLSTM, C3 - BiLSTM with word embeddings extended using polarity dictionary, C4 - BERT. Evaluation types are explained in Section 5.

Type	Classifier	SP	WP	0	WN	SN	AMB	F1	micro	macro
SDS-H	C1	85.64	00.00	77.54	00.00	83.59	16.44	81.60	94.39	65.19
	C2	86.53	00.00	82.15	00.00	**88.73**	31.71	85.20	**97.88**	**70.58**
	C3	**88.89**	00.00	**82.58**	00.00	88.09	35.71	**85.91**	97.54	69.46
	C4	87.66	00.00	82.47	00.00	87.99	**42.86**	85.39	97.26	70.32
SDS-M	C1	70.68	00.00	76.36	00.00	70.14	15.38	71.85	89.45	63.76
	C2	**78.01**	00.00	80.35	00.00	**75.30**	07.14	77.19	**95.54**	74.21
	C3	72.86	**18.18**	78.88	00.00	74.66	**25.64**	75.06	94.76	72.98
	C4	76.79	00.00	**81.08**	00.00	75.25	00.00	**77.33**	94.99	**74.76**
DOS-H	C1	60.11	00.00	48.83	00.00	61.30	00.00	55.53	**89.46**	63.81
	C2	**69.56**	00.00	54.45	00.00	67.98	06.15	62.87	87.99	**70.85**
	C3	72.05	00.00	**56.16**	00.00	**68.98**	**06.35**	**64.93**	88.48	69.50
	C4	65.46	00.00	51.74	00.00	60.85	00.00	58.04	86.23	68.04
DOS-M	C1	53.08	00.00	63.29	00.00	64.45	20.69	61.19	**94.49**	**65.29**
	C2	58.30	00.00	**67.40**	00.00	**69.37**	**37.84**	**65.98**	90.43	64.14
	C3	59.54	00.00	66.11	00.00	68.58	20.69	65.28	89.50	60.60
	C4	**61.10**	00.00	65.95	00.00	67.84	16.67	65.09	89.35	59.50
MDS	C1	77.14	00.00	76.43	00.00	76.06	17.82	75.25	89.60	61.60
	C2	**84.37**	00.00	82.30	00.00	82.78	**35.09**	82.11	96.66	**75.11**
	C3	76.00	00.00	77.29	00.00	76.54	16.22	75.42	94.98	70.77
	C4	84.14	00.00	**83.39**	00.00	**83.52**	25.50	**82.28**	**96.83**	74.25

Table 5: F1-scores for sentence-oriented evaluation. Training sets for evaluation types are the same as in Table 2 rows 10-14. Classifiers: C1 - fastText, C2 - BiLSTM, C3 - BiLSTM with word embeddings extended using polarity dictionary, C4 - BERT. Evaluation types are explained in Section 5.

BERT's performance is below the expectations of this advanced method. Looking at both tables (4 and 5), BERT's results are the best in 19 out of 69 label-specific cases, which is exactly as many as fastText was. BiLSTM outperformed other methods in 31 cases. Adding an external sentiment dictionary helped only in 14 label-specific cases. BERT dominance is observed in DOT and MDT cases, especially when analysing general metric values, where the predominance of the method is visible in 11 out of 15 cases. The advantage is repeated for MDS but not for DOS. MDT case is the most promising in terms of the further use of the recognition method in applications such as brand monitoring or early crisis detection. Figure 2 shows the ROC curves (Meistrell, 1990) for this case. The values of the general F1, micro AUC and macro AUC are the highest for the BERT method (see Table 2).

We plan to publish the data created as part of the presented works on an open license soon. We also intend to test the contextualized embedding that we are currently building using the ELMo deep word representations method (Peters et al., 2018), with the use of the large KGR10 corpus presented in work (Kocoń et al., 2019). We also want to train the basic BERT model with the use of KGR10 to investigate whether it will improve the quality of sentiment recognition. It is also very interesting to use the propagation of sentiment annotation in WordNet (Kocoń et al., 2018a,b), to increase the coverage of the sentiment dictionary and to potentially improve the recognition quality as well. This objective can be achieved by other complex methods such as OpenAI GPT-2 (Radford et al., 2019) and domain dictionaries construction methods utilising WordNet (Kocoń and Marcińczuk, 2016).

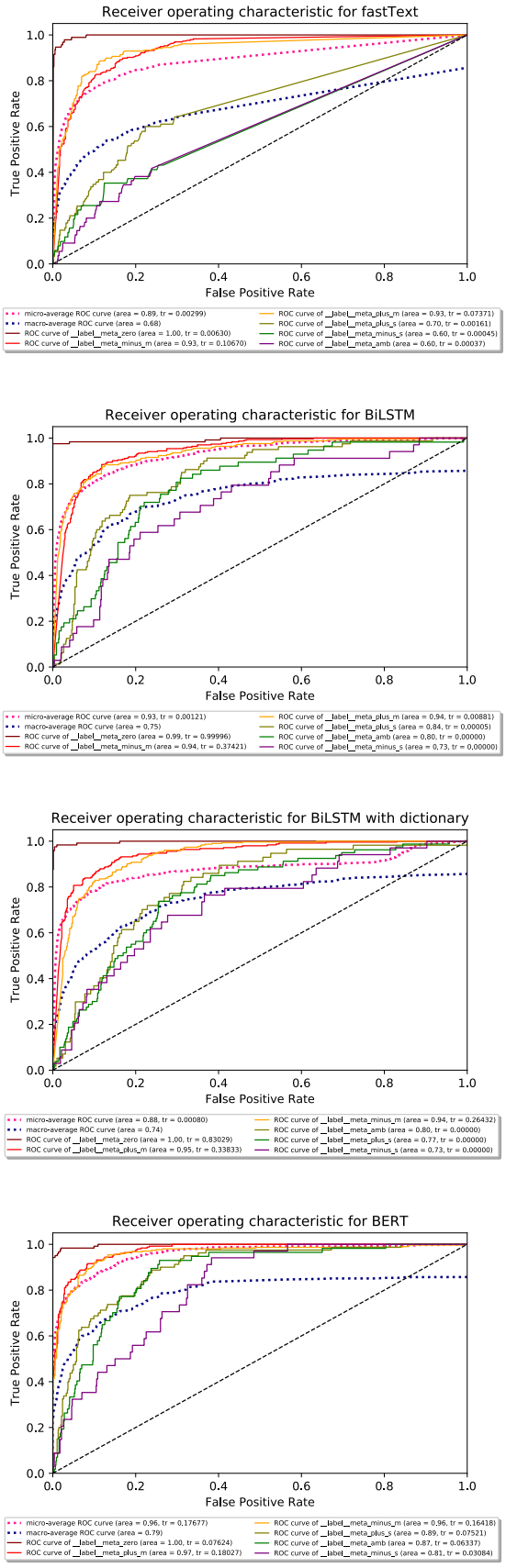

Figure 2: Receiver Operating Characteristic curves of all used classifiers for Mixed Domain Text setting.

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*[8].

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: an enhanced lexical

[8] http://w3a.pl/projekty/

resource for sentiment analysis and opinion mining. In *LREC*. volume 10, pages 2200–2204.

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

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

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*. pages 513–520.

Emitza Guzman and Walid Maalej. 2014. How do users like this feature? a fine grained sentiment analysis of app reviews. In *2014 IEEE 22nd international requirements engineering conference (RE)*. IEEE, pages 153–162.

George Hripcsak and Adam S. Rothschild. 2005. Technical Brief: Agreement, the F-Measure, and Reliability in Information Retrieval. *JAMIA* 12(3):296–298. https://doi.org/10.1197/jamia.M1733.

Arkadiusz Janz, Jan Kocoń, Maciej Piasecki, and Monika Zaśko-Zielińska. 2017. plWordNet as a Basis for Large Emotive Lexicons of Polish. In *LTC'17 8th Language and Technology Conference*. Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu, Poznań, Poland.

Armand Joulin, 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*. Association for Computational Linguistics, pages 427–431.

Jaap Kamps, Maarten Marx, Robert J Mokken, Maarten De Rijke, et al. 2004. Using wordnet to measure semantic orientations of adjectives. In *LREC*. Citeseer, volume 4, pages 1115–1118.

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

Jan Kocoń and Michał Gawor. 2018. Evaluating KGR10 Polish word embeddings in the recognition of temporal expressions using BiLSTM-CRF. *Schedae Informaticae* 27. https://doi.org/10.4467/20838476SI.18.008.10413.

Jan Kocoń, 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)*.

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

Jan Kocoń, Arkadiusz Janz, Miłkowski Piotr, Monika Riegel, Małgorzata Wierzba, Artur Marchewka, Agnieszka Czoska, Damian Grimling, Barbara Konat, Konrad Juszczyk, Katarzyna Klessa, and Maciej Piasecki. 2019. Recognition of emotions, polarity and arousal in large-scale multi-domain text reviews. In Zygmunt Vetulani and Patrick Paroubek, editors, *Human Language Technologies as a Challenge for Computer Science and Linguistics*, Wydawnictwo Nauka i Innowacje, Poznań, Poland, pages 274–280.

Jan Kocoń 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*. Springer, Brno, Czech Republic, volume 9924 of *Lecture Notes in Artificial Intelligence*.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse* 8(3):243–281.

Michael L Meistrell. 1990. Evaluation of neural network performance by receiver operating characteristic (roc) analysis: examples from the biotechnology domain. *Computer Methods and Programs in Biomedicine* 32(1):73–80.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*. volume 10, pages 1320–1326.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, page 271.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, pages 2227–2237. https://doi.org/10.18653/v1/N18-1202.

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

Adam Radziszewski. 2013. A tiered CRF tagger for Polish. In *Intelligent tools for building a scientific information platform*, Springer, pages 215–230.

Maite Taboada, Kimberly Voll, and Julian Brooke. 2008. Extracting sentiment as a function of discourse structure and topicality. *Simon Fraser Univeristy School of Computing Science Technical Report* .

Hiroya Takamura, Takashi Inui, and Manabu Okumura. 2005. Extracting semantic orientations of words using spin model. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, pages 133–140.

Amos Tversky 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.

Byron C Wallace, Eugene Charniak, et al. 2015. Sparse, contextually informed models for irony detection: Exploiting user communities, entities and sentiment. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. volume 1, pages 1035–1044.

Fei Wang, Yunfang Wu, and Likun Qiu. 2012. Exploiting discourse relations for sentiment analysis. *Proceedings of COLING 2012: Posters* pages 1311–1320.

Sida Wang and Christopher D Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers-volume 2*. Association for Computational Linguistics, pages 90–94.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Monika Zaśko-Zielińska, 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*. pages 721–730.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. volume 2, pages 207–212.