electronics

MDPI

Article

Real-Time Sentiment Analysis for Polish Dialog Systems Using MT as Pivot

Krzysztof Wołk

Polish-Japanese Academy of Information Technology, 02-008 Warsaw, Poland; kwolk@pja.edu.pl

Abstract: We live in a time when dialogue systems are becoming a very popular tool. It is estimated that in 2021 more than 80% of communication with customers on the first line of service will be based on chatbots. They enter not only the retail market but also various other industries, e.g., they are used for medical interviews, information gathering or preliminary assessment and classification of problems. Unfortunately, when these work incorrectly it leads to dissatisfaction. Such systems have the possibility of contacting a human consultant with a special command, but this is not the point. The dialog system should provide a good, uninterrupted and fluid experience and not show that it is an artificial creation. Analysing the sentiment of the entire dialogue in real time can provide a solution to this problem. In our study, we focus on studying the methods of analysing the sentiment of dialogues based on machine learning for the English language and the morphologically complex Polish language, which also represents a language with a small amount of training resources. We analyse the methods directly and use the machine translator as an intermediary, thus checking the quality changes between models based on limited resources and those based on much larger English but machine translated texts. We manage to obtain over 89% accuracy using BERT-based models. We make recommendations in this regard, also taking into account the cost aspect of implementing and maintaining such a system.

Keywords: sentiment analysis; polish sentiment; machine learning; machine translation; dialog systems; dialog sentiment; sentiment based user satisfaction

Citation: Wołk, K. Real-Time Sentiment Analysis for Polish Dialog Systems Using MT as Pivot. *Electronics* **2021**, *10*, 1813. https://doi.org/10.3390/electronics10151813

Academic Editors: Diego Reforgiato Recupero, Harald Sack and Danilo Dessì

Received: 9 June 2021
Accepted: 26 July 2021
Published: 28 July 2021

1. Introduction

Chatbots are used in many service industries to answer customer questions and help them navigate the company's website. Due to them, customers can continue to engage in the life of the company. Chatbots are expected to be a constant trend in meeting these expectations.

Currently, dialog systems are used in many areas of industry and entertainment. They ceased to be simple gadgets that with some probability would be able to interpret questions asked in natural language through keywords and answer questions based on the FAQ and they became sophisticated tools based on artificial intelligence [1,2]. Currently, deep dialogue systems analyse the grammar, syntax and meaning of natural language, which enables them to accurately interpret human utterances. Their precision of operation is so great that many industries on their first line of technical support offer chatbots [3]. Algorithms based on the so-called deep machine learning often have the ability to spontaneously execute commands of various types, e.g., turning various services on and off, etc., without human participation or verification [4].

According to [5], the growing popularity of on-demand instant messaging has changed consumer preferences in terms of communication. More and more industries are incorporating chatbots into their business process. Bots are a critical resource for improving the consumer service. Chatbots are changing the way companies communicate with

current and potential customers. Finance, healthcare, education, travel and the real estate industry derive the greatest profits from chatbots. It is predicted that 80% of companies will integrate some form of chatbot system in 2021, which will allow companies to save up to 30% of customer support costs. It turns out that over 50% of customers predict that companies are open 24/7, especially those providing international offers.

As communication technology advances, consumers expect to find information or contact customer support quickly and easily. Failure to respond promptly usually causes customers to become frustrated, which can mean losing a customer. However, contact with a human consultant is not always preferred. It turns out that 69% of consumers prefer chatbots because of their ability to provide quick answers to simple questions, 56% of consumers prefer to send a message to the company for help than call the customer service department, 37% of consumers expect quick answers in emergencies and 33% of consumers would like to use chatbots for booking, online ordering and other functions [5], which, with the current state of technology, is no longer a wishful thinking but a feasible task [6].

Chatbots are expected to become more human, but are unable to process the client's intentions, leading to misinterpreted requests and responses. They lack conversational intelligence—that is, they often fail to process the nuances implied in dialogue, resulting in inadequate conversation. However, the goal is for chatbots to be able to provide a personalized experience unique to each customer to build positive relationships, increase customer loyalty and earn positive feedback. To make it possible, they integrate various other systems, such as automatic speech recognition (ASR) [2], which are to improve this communication. Additionally, dialog systems are often integrated with a machine translator to reduce costs, which significantly increases the reach of the bot [7,8].

Unfortunately, not only chatbots but also other machine learning-based systems do not work 100% correctly and are at risk of error. The more systems connected, the greater the risk. Among others, therefore, neural dialog systems sometimes generate short and nonsensical responses and sometimes loop or fail to interpret queries correctly. Part of our project is a real-time speech analysis module, which its task is to assess the user's mood on a current basis through sentiment analysis. It analyses the flow of the entire conversation and augments the results each time new users input is provided. In this respect, we implement a dialog system, machine translator, ASR module and the results of these modules are subjected to sentiment analysis. We perform sentiment analysis using various methods for Polish and English and we also analyse the impact of machine translation on the quality of sentiment analysis. We perform analyses and evaluations based on human evaluation of the operation of our methods.

Detecting a drop in satisfaction in communication with a dialog agent is a very important problem for improving customer satisfaction. A quick response will, on the one hand, redirect the interlocutor to a human agent, on the other hand, it will allow us to collect data on in which cases the agent fails and remove these problems.

The article is structured as follows. Section 2 discusses the current state of the art as far as sentiment analysis is concerned. It also describes the experimental environment focusing machine translation, ASR, TTS and dialog system and their connection in a pipeline of tools. In Section 3 experiments are conducted and divided in subsections by the language of the trained models. Section 4 provides manual confirmation of the results obtained in Section 3. Finally, in Section 5 we draw and discuss conclusions.

2. Experimental Environment and the Current State of Knowledge

The aim of the study was to implement an analytical module to, among other reasons, analyse the sentiment of the text coming from dialogues between man and machine. This module was designed to analyse dialogue in real time and react to increased user dissatisfaction by redirecting it to a human agent. For this purpose, a test environment consisting of dialog system modules was prepared and the analysis of sentiment was carried out based on their results.

2.1. Testing Environment

Our dialog system consisted of a dialog agent implemented in English using DeepPavlov [9] and a model trained in the BERT architecture [10]. For training the model we used transcriptions of real life problems and transfer learning of the English model within the DeepPavlov toolkit. The Polish language was handled by a machine translator based on convolutional neural networks implemented as part of the ModernMT tool [11], where, apart from our own small body, we used a corpus based on subtitles [12]. For the purposes of the ASR module, the KALDI tool [13] was used and all audio recordings and their transcriptions produced under the Clarin project [14]. The scheme of the system operation is presented in Figure 1.

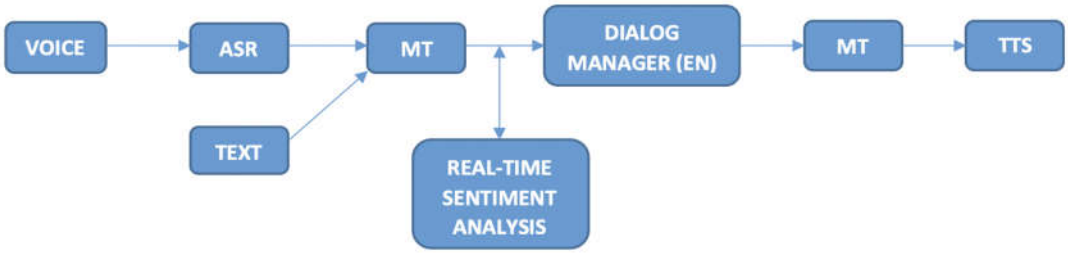

Figure 1. Diagram of the dialog system with the analysis of the dialog sentiment.

Each of the modules was evaluated with a separate metric. The dialog system was verified with the SQuAD metric [15], reaching 81.5% and 93.4% on our own evaluation corpus (containing 50 questions and answers consisting of 1672 sentences). The BLEU metric [16] was used to assess machine translation (MT) and for the PL to EN translation, 64.21 points were achieved and 43.23 points in the reverse direction. On the other hand, the ASR system for PL was assessed using the WER metric [17], obtaining 18.21 and for EN 12.37. The MT and ASR modules were assessed on the aforementioned 1672 sentences prepared and translated by humans as part of our work. These results prove their high quality, consistent with the current state of knowledge in these fields, which should not disturb the reliability of the results of the sentiment analysis based on their results.

2.2. Sentiment Analysis Techniques Used

As part of the sentiment analysis itself, we made an in-depth analysis of the current state of knowledge and selected the most popular and most promising approaches. We started the work with the initial verification of the popular Vader tool [18], the results of which we treated as a reference point for further experiments.

The Vader tool uses a rule-based model using an English-language lexicon prepared by the authors [19]. It detects denials, words that enhance the overall tone and even pays attention to the case of letters or the number of exclamation points. The authors of the tool have prepared a lexicon with particular emphasis on the social media vocabulary, including emoticons and slang. The tool does not follow a machine learning approach, has consistent sentence grades and cannot be trained. Only English is supported. In order to analyse the text in Polish, each query had to be machine translated. A poorly translated text could have resulted in an inaccurate analysis. Since the BLEU score of our MT system was over 60 points, in our opinion translation quality was satisfactory for the experiment. According to [20] the impact of MT quality should be marginal. Additionally, not all rules contained in the tool translate to all languages. Therefore, there are adaptations of the tool, e.g., the German-language GerVADER [21]. The Vader tool was created in 2014 and at that time it was better than other methods [22].

After experiments with the VADER tool, sentiment analyses based on supervised learning methods were added [23]. More specifically, the Linear SCV [24] and naive Bayes

Bernoulli [25] methods available in the sci-kit learn library [26] were tested. For their training, the English-language data set Sentiment140 [27] was selected, containing entries from the social network Twitter along with sentiment markers. Models were trained on the entire dataset using TF–IDF text vectorization. Additionally, for all methods, the returned results were normalized so that the values for the positive, neutral and negative sentiment represent the binary marking—0 for the negative sentiment and 1 for the positive sentiment. Our implementation was adjusted to show both the results of 2-class (positive and negative sentiment), 3-class (positive, neutral and negative sentiment) and percentages.

The Polish training set was created from approximately 43,000 tagged entries, also from Twitter [28]. Entries were downloaded through the API, using a simple script using the tweepy library [29]. On the basis of the prepared Polish data set, 2-class LinearSVC and naive Bayes Bernoulli models were trained. The Polish dataset was also used to train the naive Bayes Bernoulli 3-class model.

After these basic models were prepared, the state-of-the-art models that performed best were analysed. The world's best models with an efficiency of over 95% were created on the basis of transfer learning techniques. Due to fine tuning, the model can be adapted to up to 20 different activities, such as the sentiment analysis, answers to open-ended questions, text classification, translation, etc. To achieve such high efficiency, data sets with sizes over 20 TB and GPU/TPU units for training and fine tuning are needed. Sets and pretrained models are available for download and the algorithms can be adapted to the equipment – here are ways to reduce models, e.g., by 60% with a slight loss of quality (by 2%) – [30] which we did. Model rankings were created due to the machine learning community [31].

On the basis of these analyses, pretrained: DistilBERT [32], T5 (text-to-text transfer-transformer) [33] and XLNET [34] were added. The transformers library by HuggingFace [35] was used for this. It supports the formats of popular libraries PyTorch [36] and TensorFlow [37]. The spaCy wrapper [38] was also created for it, which simplifies the process of model training. The transformers library has the ability to export models to the onnx format, which in turn will allow the model to be optimized for the production environment [39].

Similar methods were already applied to other Slavic languages, but to other topic domains. Authors of [40] apply the sentiment analysis to the financial context news in the Lithuanian language. In [41] authors apply a modified RoBERTa model for sentiment analysis in Czech, which they find to be most successful.

The ABSA analysis tool (aspect target sentiment analysis) was also used [42]. It works by examining the sentiment of the selected subject (aspect) in the text. Due to this approach, from the user's opinion, one can obtain information about what exactly is considered good in the product and what is bad, e.g., from the opinion "I like my phone, the camera works great, but the battery leaves a lot to be desired" we get the analysis result: aspect: camera—sentiment: positive and aspect: battery—sentiment: negative. A proof-of concept was prepared, which uses the DistilBERT model adapted to such an analysis.

3. Experiments

One of the first tasks was to compare two text vectorization methods (Table 1), for two different models—naive Bayes Bernoulli and LinearSVC on the Amazon Video Games set [43]. The TF–IDF method [44] has been compiled with an implementation called LabelEncoder [45], which marks each word with a number.

Metrics that were included are: precision, recall and F1 score [46].

Table 1. Comparison of text vectorization methods.

LinearSVC	Precision	Recall	F1 Score
Word Embedding	0.885928	0.914875	0.900169
TF–IDF	0.914081	0.987962	0.949587

Naive Bayes Bernoulli			
Word Embedding	0.965066	0.376812	0.541999
TF–IDF	0.906364	0.849957	0.877255

TF–IDF in the case of LinearSVC gave slightly better results. In the case of the naive Bayes model, it can be seen that word embedding gave 6% more precision. Nevertheless, the ratio of correct observations to the entire recall was twice lower for this model, which can also be seen from the F1 result, whose recall is a component.

3.1. Pretrained English Models

Models that arose after significant development in the field of transfer learning, i.e., transformer models, are pretrained, which also means that they already have mechanisms for converting text into vectors in hidden layers integrated with the model. Examples of such models are, e.g., BERT or GPT [47].

In this respect, the DistilBERT model was trained on the SST-2 benchmark (Stanford Sentiment Treebank v2) [48]. The transformers library was used to train the DistilBERT model and the SST-2 task came from the GLUE Benchmark [49] set, the result was similar to that recorded in the model ranking, i.e., accuracy (eval_acc) around 0.92:

eval_loss = 0.3662;

eval_acc = 0.9013;

epoch = 3.0.

The effectiveness of pretrained English-language models was also compared, as presented in Table 2. In preparation for measuring and comparing the model results, a simple script was prepared. The Amazon—"Video Games" in English [50] was selected to evaluate the effectiveness. The dataset is in the form of opinions rated 1–5. Ratings 1–2 were negative, 3 neutral and 4–5 positive.

The naive Bayes and LinearSVC models were trained on the entire Sentiment140 dataset using TF–IDF vectorization. DistilBERT was trained on Wikipedia + BookCorpus corpora. The T5 model was pretrained on the C4 corpus [51] and the XLNET model on BookCorpus, English Wikipedia, Giga5, ClueWeb and CommonCrawl [52]. The Vader model was also used for the compilation, based on a set of rules and a specially prepared corpus. The evaluation was based on 20,000 records.

Table 2. Comparison of the effectiveness of the models.

XLNET	**T5**	**DistilBERT**	**Naive Bayes Bernoulli**	**LinearSVC**	**Vader**
31.09%	79.99%	84.73%	65.78%	75.69%	84.70%

It has been noticed that the T5 model sporadically generates an unexpected sequence (e.g., "Sst" instead of prediction), which, despite a high score, indicates incorrect pretraining of this network. The problem was not investigated further as the Polish-language models were the priority in the study.

3.2. Comparative Experiments for Polish-Language Models

In terms of the Polish language, the Polish RoBERTa [53] and PolBERT [54] models were trained on various corpora with binary sentiment markings and the metrics of effectiveness were recorded. The simpletransformers library was used for this purpose [55]. The results of the experiments are presented in Table 3.

Table 3. Results of comparative experiments for Polish-language models.

	Clarin	PolElmo 2.0	Allegro Reviews	Result
Polish RoBERTa	-	-	-	acc = 0.5751 eval_loss = 0.6870 f1 = 0.0839 mcc = -0.0172
Polish RoBERTa	x	-	-	acc = 0.7790 eval_loss = 0.5570 f1 = 0.8060 mcc = 0.5510
Polish RoBERTa	-	x	-	acc = 0.8625 eval_loss = 0.3680 f1 = 0.8358 mcc = 0.7177
Polish RoBERTa	-	-	x	acc = 0.7936 eval_loss = 0.5181 f1 = 0.8849 mcc = 0.0
Polish RoBERTa	x	x	-	acc = 0.8930 eval_loss = 0.3228 f1 = 0.8685 mcc = 0.7786
Polish RoBERTa	-	x	x	acc = 0.5864 eval_loss = 0.6782 f1 = 0.0 mcc = 0.0
PolBERT	-	-	-	acc = 0.5965 eval_loss = 0.6772 f1 = 0.7038 mcc = 0.1110
PolBERT	-	x	-	acc = 0.8804 eval_loss = 0.3225 f1 = 0.8538 mcc = 0.7528
PolBERT	x	-	-	acc = 0.7961 eval_loss = 0.4759 f1 = 0.8366 mcc = 0.5718
PolBERT	x	x	-	acc = 0.7913 eval_loss = 0.6397 f1 = 0.8255 mcc = 0.5660

The data sets used were:

- Clarin—Polish entries on the social platform Tweeter [56];
- PolEmo 2.0—Multidomain product review [57];
- AllegroReviews—Multidomain product reviews [58].

Models obtained the best results (0.88 and 0.89) after training on the PolEmo 2.0 corpus. It is worth paying attention to training RoBERTa on Allegro Reviews and then on

PolEmo 2.0—the accuracy obtained was 0.58. Therefore, the domain of the corpora is important—PolEmo mainly consists of opinions about places and Allegro Reviews about products. If the subject matter overlaps, and this is partly the case with Clarin and PolEmo, the result should be better, due to the uniform context of the statements made, among other reasons.

The MCC metric was also included [59]. A result close to 1 represents a perfect prediction, 0 is no better than a random prediction and −1 represents a complete mismatch. In two cases, the MCC metric was equal to 0—this is most likely an error on the side of the used library.

3.3. Possible Optimization of Models

Optimizations play a key role as they can accelerate the model by up to 30%, thus reducing operating costs. Model inference, or otherwise obtaining the prediction result, can be optimized using tools such as the OpenVINO [60] framework, ONNX of the Microsoft company [61] or TensorRT [62]. KITO is also available [63], which is mainly used for image processing models. An extensive article by Intel on system, application and model optimization explains the importance of optimizations at the level of the entire infrastructure [64]. Finally, it was necessary to check the tools and additionally apply model pruning [65].

3.4. Results of Polish Models

During the experiments based on Polish models, the sentiment for 10,000 sentences from the Clarin set was analysed. The results are shown in Table 4.

Table 4. Results of experiments for Polish-language models.

Model	AVG CPU%	MAX CPU%	AVG MEM	MAX MEM	Inference time	Accuracy%
LinearSVC_PL	385.9	1057.9	162.3 MB	315.4 MB	0.001117994379997	98.77
PolishRoberta	1548.81	4276.68	2.5 GB	4.3 GB	2.8007734849453	93.07
Polbert	1123.87	4107.96	4.2 GB	6.7 GB	2.48969090027809	92.97
NaiveBayes_2_CLS_PL	161.33	1141.81	309.3 MB	322.1 MB	0.012122882723808	89.38
NaiveBayes_3_CLS_PL	146.47	1028.85	314.4 MB	325.3 MB	0.014599187135696	88.31
LinearSVC_EN	99.95	100.22	1.1 GB	1.3 GB	0.003802319741249	87.27
NaiveBayes_2_CLS_EN	109.67	1045.76	1.3 GB	1.4 GB	0.072344211030006	83.29
T5	3936.15	4604.82	3.0 GB	3.0 GB	0.640251659822464	75.36
DistilBERT	3498.85	4106.75	2.6 GB	3.0 GB	0.356072693634033	70.88
Vader_2_CLS_EN	100.02	100.69	1.1 GB	1.2 GB	0.009924677634239	65.87
Vader_3_CLS_EN	564.29	1028.49	610.2 MB	1.1 GB	8.09271097183228E-05	60.38
Vader_PERCENTAGE_EN	582.74	1065.38	607.8 MB	1.1 GB	8.22359085083008E-05	58.13
XLNET	3850.74	4490.56	3.0 GB	3.3 GB	0.501096723675728	44.04

The results of PolBert and PolishRoberta are similar to those of the Polish ranking of models and presented in the Table 5 [49]:

Table 5. The results of experiments for the Polish-language models.

Model	AVG CPU%	MAX CPU%	AVG MEM	MAX MEM	Inference time	Accuracy%
Polbert	1032.09	4081.53	4.4 GB	6.8 GB	2.57093233888149	92.95
PolishRoberta	1542.35	4358.97	2.6 GB	4.7 GB	2.87089015738964	92.5
LinearSVC_PL	268.9	644.87	162.8 MB	301.6 MB	0.001157759809494	90.58
NaiveBayes_2_CLS_PL	125.06	525.94	299.6 MB	307.8 MB	0.01218701300621	84.05
NaiveBayes_3_CLS_PL	125.84	644.87	295.1 MB	309.5 MB	0.014857436347008	77.2

4. Manual Evaluation

Due to the fact that the results presented in Section 3 prove that not only the general metric is very important for the quality of the evaluation, but also the area in which the products under analysis move, it was also decided that we would conduct a manual analysis. Although our models achieved some metric of popular benchmarks at the world level, due to the fact that some of them used the transfer learning of models belonging to other domains, they could potentially not be such good predictors in our field.

Therefore, 100 real opinions in Polish and 100 in English from our clients were manually prepared. They were subjected to manual human evaluation and then compared using the methods described. Manual evaluation of the sentiment of each of the comments was made using the following scale:

"-"—negative sentiment of the comment;
"0"—neutral sentiment of the comment;
"+"—positive sentiment of the comment.

During the test, points were awarded for compliance with the subjective assessment of the subject. The method received 3 points for perfect compliance and for partial compliance (e.g., the method considers the comment as neutral, the user as positive), 1 point. No compliance resulted in 0 points. The analysis was performed separately for 2-class models and separately for 3-class models. Finally, for each method, the percentage of compliance with the human method was determined.

The results of the human evaluation are presented in the form of graphs. Figure 2 shows the test result of the sentiment testing methods for Polish comments. Figure 3 shows the result of the test of sentiment research methods for Polish comments that are not considered neutral. Figure 4 shows the test results of the sentiment test method for English comments and Figure 5 shows the test results of sentiment test methods for English comments without those considered neutral.

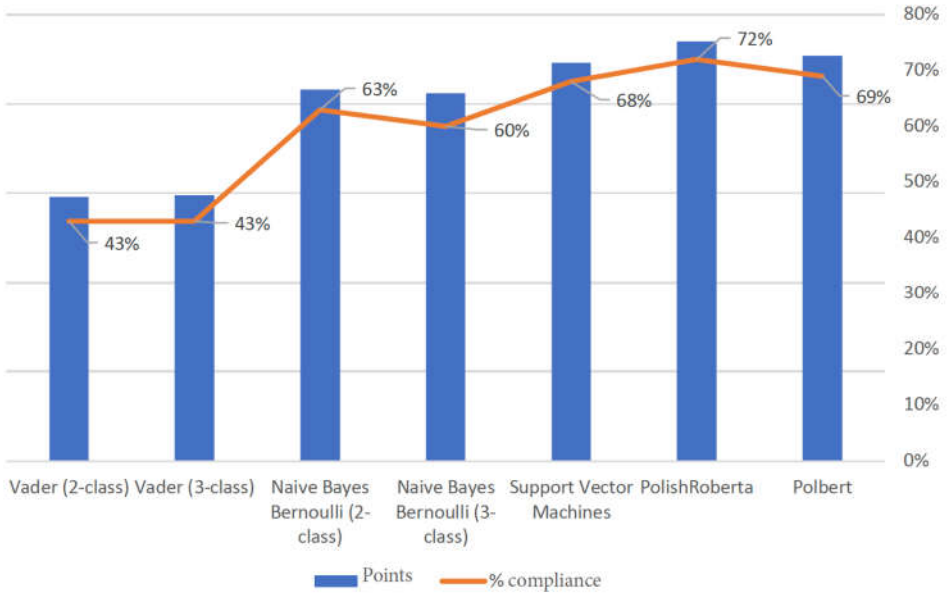

Figure 2. Test of sentiment research methods for Polish comments.

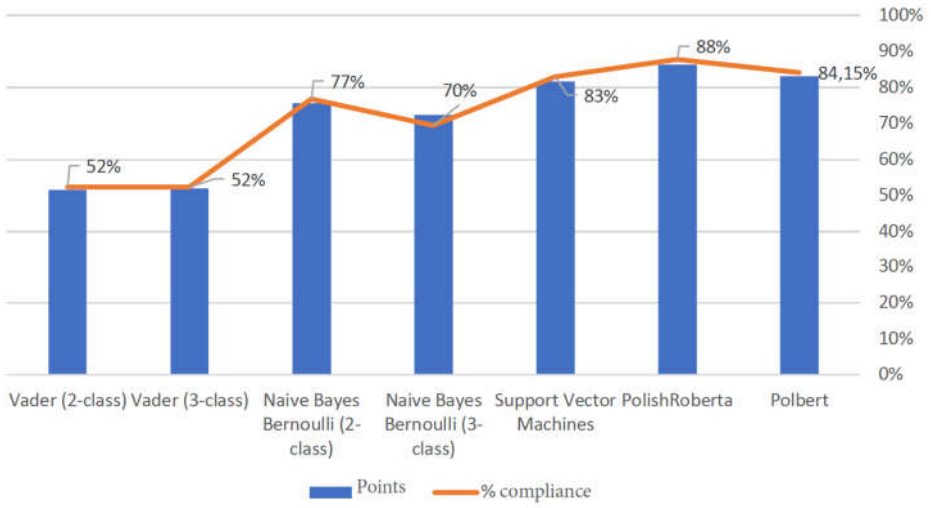

Figure 3. Test of sentiment research methods for Polish comments without those considered neutral.

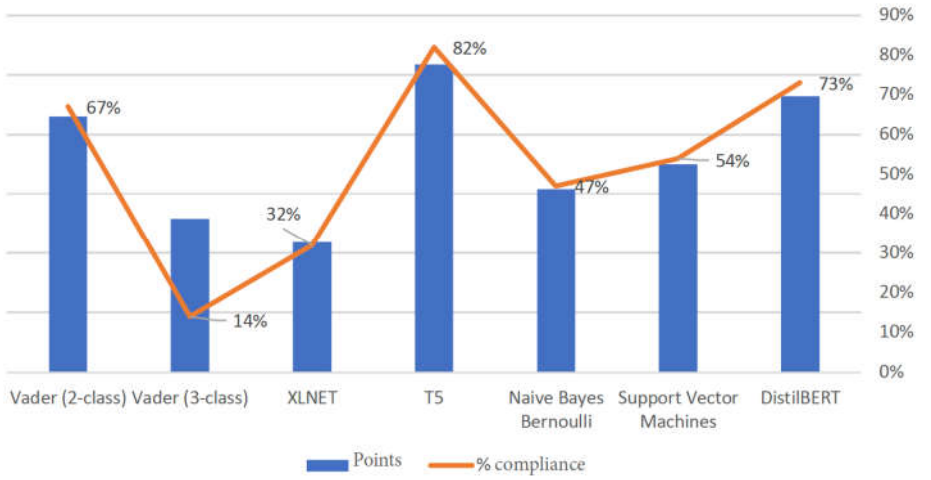

Figure 4. Test of sentiment test methods for English comments.

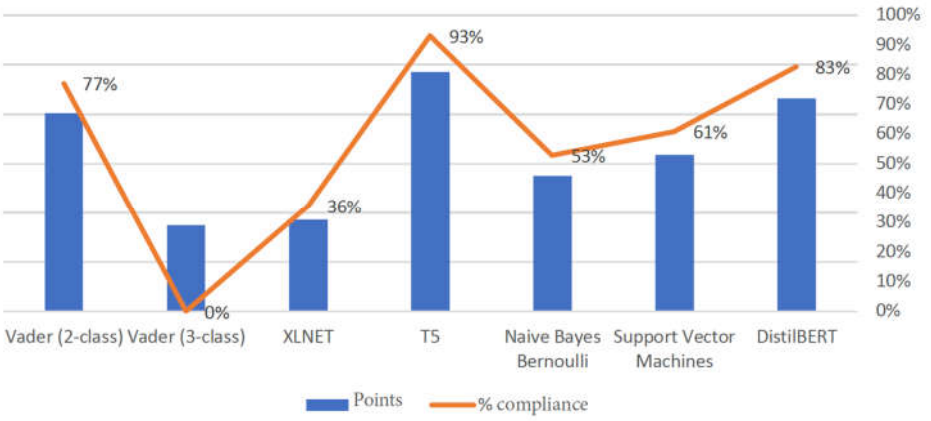

Figure 5. Test of sentiment test methods for English comments without those considered neutral.

5. Conclusions and Discussion

Seemingly, due to the fact that the predictions are to be made in real time, the main determinant of accuracy will be the inference time. This will allow for cost optimization on the part of the enterprise.

BERT-based models are pretrained on large datasets, so their domain adaptability should be high. The problem of scaling, however, is the inference time and the use of resources by such a model. The inference time can be reduced by optimization methods, e.g., by exporting models to the ONNX format, using distillation or pruning.

The seemingly proposed optimal solution based on generic benchmarks could be the implementation of a queuing system for prediction, based on the BERT—Polbert and DistilBert models. If the queue was full, the models would be supported by the less demanding LinearSVC, Vader and naive Bayes. With large discrepancy in inference time between models and with a large number of queries, the supporting models will provide more results.

However, manual analysis revealed that for the Polish language, the PolishRoberta method turned out to be the most consistent. For the English language, the T5 method turned out to be the most compatible. However, the Vader method was well below expectations. It was with this method in mind that we tried to machine translate queries to avoid the need to create our own rules adapted to the language. It turns out, however, that the method not only fares poorly, but machine translations, and in particular ASR, significantly worsen its results. This is most likely because we lose a lot of information due to the normalization of the text on which the rules used in it are based on.

In conclusion, we reviewed the sentiment analysis techniques that achieve the highest results on generic benchmarks and we checked which of them in real business use work best in terms of quality and how they scale in performance. This is valuable knowledge from the business and implementation point of view. There is no doubt, however, that it is possible to further develop the research towards the analysis of model domain adaptation techniques and optimization of their performance. In a company, even a few percent of yields in these areas on a macroscale translate into real money.

Funding: This work was funded by The National Centre for Research and Development in Poland, grant agreement number POIR.01.01.01-00-0009/19.

Conflicts of Interest: The authors declare no conflict of interest.

Reference

1. Ahmed, A.; Ali, N.; Aziz, S.; Abd-alrazaq, A.A.; Hassan, A.; Khalifa, M.; Elhusein, B.; Ahmed, M.; Ahmed, M.A.S.; Househ, M. A review of mobile chatbot apps for anxiety and depression and their self-care features. *Comput. Methods Programs Biomed. Update* **2021**, 100012.
2. Cahn, J. CHATBOT: Architecture, Design, & Development. Senior Thesis, University of Pennsylvania, Philadelphia, PA, USA, 2017.
3. Xu, A.; Liu, Z.; Guo, Y.; Sinha, V.; Akkiraju, R. A new chatbot for customer service on social media. In Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, Denver CO USA, 6–11 May 2017; pp. 3506–3510.
4. Csaky, R. Deep learning based chatbot models. *arXiv* **2019**, arXiv:1908.08835.
5. Linchpin, T. 25 Chatbot Stats and Trends Shaping Businesses in 2021. Avalilable online: https://linchpinseo.com/chatbot-statistics-trends/ (accessed on 28 May 2021)
6. Singh, S.; Thakur, H.K. Survey of Various AI Chatbots Based on Technology Used. In Proceedings of the 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 4–5 June 2020; pp. 1074–1079.
7. Nosotti, E. Building a Multi-Language Chatbot with Automated Translations. Available online: https://medium.com/rockedscience/building-a-multi-language-chatbot-with-automated-translations-e2acd053bc5c (accessed 29 05 2021).
8. Hu, W.; Le, R.; Liu, B.; Ma, J.; Zhao, D.; Yan, R. Translation vs. Dialogue: A Comparative Analysis of Sequence-to-Sequence Modeling. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 4111–4122.
9. Soloveva, A. SO at SemEval-2020 task 7: DeepPavlov logistic regression with BERT embeddings vs. SVR at funniness evaluation. In Proceedings of the Fourteenth Workshop on Semantic Evaluation, Barcelona, Spain, 12–13 December 2020; pp. 1055–1059.

10. Rana, M. Eaglebot: A Chatbot Based Multi-Tier Question Answering System For Retrieving Answers From Heterogeneous Sources Using BERT. Master Thesis, Georgia Southern University, Statesboro, GA, USA, 2019; pp. 431–437.
11. Germann, U.; Barbu, E.; Bentivogli, L.; Bertoldi, N.; Bogoychev, N.; Buck, C.; van der Meer, J. Modern MT: A new open-source machine translation platform for the translation industry. Baltic J. Mod. Comput. **2016**, *4*, 397.
12. Tiedemann, J. Parallel Data, Tools and Interfaces in OPUS. In Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012), Istanbul, Turkey, 23–25 May 2012; pp. 2214–2218.
13. Guglani, J.; Mishra, A.N. Continuous Punjabi speech recognition model based on Kaldi ASR toolkit. *Int. J. Speech Technol.* **2018**, *21*, 211–216.
14. Draxler, C.; van den Heuvel, H.; van Hessen, A.; Calamai, S.; Corti, L. A CLARIN Transcription Portal for Interview Data. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 13–15 May 2020; pp. 3353–3359.
15. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv* **2016**, arXiv:1606.05250.
16. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; pp. 311–318.
17. Favre, B.; Cheung, K.; Kazemian, S.; Lee, A.; Liu, Y.; Munteanu, C.; Zeller, F. Automatic human utility evaluation of ASR systems: Does WER really predict performance? In Proceedings of the INTERSPEECH 2013 14thAnnual Conference of the International Speech Communication Association, 25–29 August 2013; pp. 3463–3467.
18. Borg, A.; Boldt, M. Using VADER sentiment and SVM for predicting customer response sentiment. *Expert Syst. Appl.* **2020**, *162*, 113746.
19. Hutto, C.; Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the International AAAI Conference on Web and Social Media, Ann Arbor, MI, USA, 1–4 June 2014; 216–225.
20. Zhang, Y.; Vogel, S.; Waibel, A. Interpreting bleu/nist scores: How much improvement do we need to have a better system? In Proceedings of the LREC, Lisbon, Portugal, 26–28 May 2004.
21. Tymann, K.; Lutz, M.; Palsbröker, P.; Gips, C. GerVADER-A German Adaptation of the VADER Sentiment Analysis Tool for Social Media Texts. In Proceedings of the LWDA, Berlin, German, September 30–2 October 2019; pp. 178–189.
22. Shelar, A.; Huang, C.Y. Sentiment analysis of twitter data. In Proceedings of the 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 12–14 December 2018; pp. 1301–1302.
23. Caruana, R.; Niculescu-Mizil, A. An empirical comparison of supervised learning algorithms. In Proceedings of the 23rd international conference on Machine learning, Pittsburgh, PA, USA, 25–29 June 2006; pp. 161–168.
24. Almatarneh, S.; Gamallo, P.; Pena, F.J.R. CiTIUS-COLE at semeval-2019 task 5: Combining linguistic features to identify hate speech against immigrants and women on multilingual tweets. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, MN, USA, 6–7 June 2019; pp. 387–390.
25. Singh, G.; Kumar, B.; Gaur, L.; Tyagi, A. Comparison between multinomial and Bernoulli naïve Bayes for text classification. In Proceedings of the 2019 International Conference on Automation, Computational and Technology Management (ICACTM), London, UK, 24-26 April 2019; pp. 593–596.
26. Scikit-Learn: Machine Learning in Python. Available online: http://scikit-learn.org/stable/index.html (accessed on 28 May 2021).
27. Sentiment140. Available online: http://help.sentiment140.com/for-students (accessed on 28 May 2021).
28. Clarin SI Repository. Available online: https://www.clarin.si/repository/xmlui/ (accessed on 28 May 2021).
29. Tweepy—An Easy-to-Use Python Library for Accessing the Twitter API. Available online: https://www.tweepy.org/ (accessed on 28 May 2021).
30. Sajjad, H.; Dalvi, F.; Durrani, N.; Nakov, P. On the Effect of Dropping Layers of Pre-Trained Transformer Models. Available online: https://arxiv.org/pdf/2004.03844.pdf, 2020 (accessed on 28 May 2021).
31. Rankings on NLP. https://nlpprogress.com/,. Available online: https://paperswithcode.com/task/sentiment-analysis/latest (accessed on 28 May 2021).
32. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
33. Pipalia, K.; Bhadja, R.; Shukla, M. Comparative Analysis of Different Transformer Based Architectures Used in Sentiment Analysis. In Proceedings of the 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART), Moradabad, India, 4-5 December 2020; pp. 411–415.
34. Banerjee, S.; Jayapal, A.; Thavareesan, S. NUIG-Shubhanker@ Dravidian-CodeMix-FIRE2020: Sentiment Analysis of Code-Mixed Dravidian text using XLNet. *arXiv* **2020**, arXiv:2010.07773.
35. Transformers—Natural Language Processing for Jax, PyTorch and TensorFlow. Available online: https://github.com/huggingface/transformers (accessed on 28 May 2021).
36. Subramanian, V. *Deep Learning with PyTorch: A Practical Approach to Building Neural Network Models Using PyTorch*; Packt Publishing Ltd.: Birmingham, UK, 2018.
37. Donadi, M. A System for Sentiment Analysis of Online-Media with TensorFlow. Ph.D. Thesis, Hochschule Für Angewandte Wissenschaften Hamburg, Hamburg, Germany, 2018.

38. Sharma, M. Polarity Detection in a Cross-Lingual Sentiment Analysis using spaCy. In Proceedings of the 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 4–5 June 2020; pp. 490–496.
39. Microsoft/Onnxruntime: Onnx Runtime: Cross-Platform, High Performance Scoring Engine for mL Models. Available online: https://github.com/microsoft/onnxruntime (accessed on 28 May 2021).
40. Štrimaitis, R.; Stefanovič; P; Ramanauskaitė; S; Slotkienė; A Financial Context News Sentiment Analysis for the Lithuanian Language. *Appl. Sci.* **2021**, *11*, 4443.
41. Straka, M.; Náplava, J.; Straková; J; Samuel, D. RobeCzech: Czech RoBERTa, a monolingual contextualized language representation model. *arXiv* 2021, arXiv:2105.11314.
42. Intellica AI. Available online: https://intellica-ai.medium.com/aspect-based-sentiment-analysis-everything-you-wanted-to-know-1be41572e238 (accessed on 28 May 2021).
43. Altun, L. A Corpus Based Study: Analysis of the Positive Reviews of Amazon. com Users. *Adv. Lang. Lit. Stud.* **2019**, *10*, 123–128.
44. Yun-tao, Z.; Ling, G.; Yong-cheng, W. An improved TF-IDF approach for text classification. *J. Zhejiang Univ. Sci. A* **2005**, *6*, 49–55.
45. Kallimani, J.S. Machine Learning Based Predictive Action on Categorical Non-Sequential Data. *Recent Adv. Comput. Sci. Commun. (Former. Recent Pat. Comput. Sci.)* **2020**, *13*, 1020–1030.
46. Wang, R.; Li, J. Bayes test of precision, recall, and f1 measure for comparison of two natural language processing models. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July-2 August 2019; pp. 4135–4145.
47. Wang, B.; Shang, L.; Lioma, C.; Jiang, X.; Yang, H., Liu, Q., Simonsen, J.G. On position embeddings in bert. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 3–7 May 2021; Volume 2, pp. 12–13.
48. Sajjad, H.; Dalvi, F.; Durrani, N.; Nakov, P. Poor Man's BERT: Smaller and Faster Transformer Models. *arXiv* **2020**, arXiv:2004.03844.
49. Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S.R. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv* **2018**, arXiv:1804.07461.
50. Amazon Video Games. Available online: http://jmcauley.ucsd.edu/data/amazon/ (accessed on 28 May 2021).
51. T5: Text-To-Text Transfer Transformer. Available online: https://github.com/google-research/text-to-text-transfer-transformer#dataset-preparation (accessed on 28 May 2021).
52. Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 5753–5763.
53. Polish RoBERTa. Available online: https://github.com/sdadas/polish-roberta (accessed on 28 May 2021).
54. Kłeczek, D. Polbert: Attacking Polish NLP Tasks with Transformers. In Proceedings of the PolEval 2020 Workshop, Warszawa, Poland, 26 October 2020; pp. 79–88.
55. Zumel, P.; Garcia, O.; Cobos, J.A.; Uceda, J. Tight magnetic coupling in multiphase interleaved converters based on simple transformers. In Proceedings of the Twentieth Annual IEEE Applied Power Electronics Conference and Exposition, 2005 APEC, Busan, Korea, 18–19 November 2005; pp. 385–391.
56. Twitter Sentiment for 15 European Languages. Available online: https://www.clarin.si/repository/xmlui/handle/11356/1054 (accessed on 28 May 2021).
57. Kocoń, J.; Miłkowski, P.; Zaśko-Zielińska, M. Multi-level sentiment analysis of PolEmo 2.0: Extended corpus of multi-domain consumer reviews. In Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Hong Kong, China, 3–4 November 2019; pp. 980–991.
58. AllegroReviews Dataset. Available online: https://github.com/allegro/klejbenchmark-allegroreviews (accessed on 28 May 2021).
59. Boughorbel, S.; Jarray, F.; El-Anbari, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS ONE* **2017**, *12*, e0177678.
60. Gorbachev, Y.; Fedorov, M.; Slavutin, I.; Tugarev, A.; Fatekhov, M.; Tarkan, Y. Openvino deep learning workbench: Comprehensive analysis and tuning of neural networks inference. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea, 27 October–2 November 2019.
61. Jin, T.; Bercea, G.T.; Le, T.D.; Chen, T.; Su, G.; Imai, H.; Eichenberger, A.E. Compiling ONNX Neural Network Models Using MLIR. *arXiv* **2020**, arXiv:2008.08272.
62. Ren, H.; Manivannan, N.; Lee, G.C.; Yu, S.; Sha, P.; Conti, T.; D'Souza, N. Improving OCT B-scan of interest inference performance using TensorRT based neural network optimization. *Investig. Ophthalmol. Vis. Sci.* **2020**, *61*, 1635–1635.
63. Shanmugamani, R. *Deep Learning for Computer Vision: Expert techniques to train advanced neural networks using TensorFlow and Keras*; Packt Publishing Ltd.: Birmingham, UK, 2018.
64. Optimization Practice of Deep Learning Inference Deployment on Intel® Processors. Available online: https://software.intel.com/content/www/us/en/develop/articles/optimization-practice-of-deep-learning-inference-deployment-on-intel-processors.html (accessed on 28 May 2021).
65. Onan, A.; Korukoğlu, S.; Bulut, H. A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification. *Inf. Process. Manag.* **2017**, *53*, 814–833.