

Targeted Sentiment Analysis for Ukrainian and Russian News Articles

Iuliia Makogon¹, Igor Samokhin²

¹*Semantrum, Kyiv, Ukraine*

²*Grammarly*

Abstract

One of the most challenging problems in sentiment analysis is an analysis of multiple targets in the same text without clear boundaries between target contexts. This problem is even harder for Ukrainian and Russian languages because of the lack of datasets and established approaches. Responding to the business needs of our company, we created the bilingual dataset, manually annotated for targeted sentiment according to strict guidelines. This dataset allowed us to fine-tune a pre-trained multilingual BERT model and improve key metrics (macro F1, F1 for negative and positive classes) over the baseline models. As a by-product, we have trained new NER models for both languages. NER and targeted sentiment models were successfully introduced into the production environment at Semantrum.

Keywords


Targeted Sentiment, Dataset Annotation, BERT

1. Introduction

Semantrum is an online media monitoring and content analysis system with a focus on PR effectiveness measurement. It analyses various media sources in Ukrainian, Russian, English, and other languages. It uses the Big Data and machine learning technology, making huge datasets structured according to different criteria in real time. For each message, the system detects entities (brands, companies, persons, etc.), sentiment, and their conjunctions. The processed data is visualized in the form of diagrams, interactive infographics, and dashboards.

In 2018, the company faced the task of replacing the lexicon-based algorithm for sentiment detection with a more advanced one. The biggest challenge was the need to determine the tone of a text relative to a particular entity, as well as the fact that the main languages of the texts in the system are Ukrainian and Russian. At that time, a certain number of resources were available for Russian, although much less than for English, but there was very little research for Ukrainian. The problem of targeted sentiment is also less researched than the document-level sentiment. Most of the available datasets are in English and are based on Twitter, product reviews, or financial news. These texts differ significantly in structure and lexicon from the texts of news articles and blogs. To train a classifier, we decided to collect and annotate a dataset that would meet our business needs.

ICTERI-2021, Vol II: Workshops, September 28 – October 2, 2021, Kherson, Ukraine

 yuliia.makohon@semantrum.net (I. Makogon); igor.samokhin@grammarly.com (I. Samokhin)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



CEUR Workshop Proceedings (CEUR-WS.org)

With 987 Ukrainian and 979 Russian annotated news articles, we used a pre-trained multilingual BERT model to determine the sentiment of entities mentioned in the text. We fine-tuned the model on our data to produce a single bilingual classifier that achieved an improvement of 25 percentage points over the baseline in macro F1. With macro F1 of 0.67, the best model is still not sufficiently accurate for non-neutral fragments of text, so further work on the dataset and the model architecture is needed.

Most of the work related to the corpus construction and experiments with the dataset occurred during 2018 and 2019. After testing, the system was approved for use in production and received positive feedback from the analytics department and customers of Semantrum. The given approach showed promising results when adopted for other European languages and multilingual sentiment analysis.

Previously, Kanishcheva and Bobichev in [1] and [2] explored the problem of sentiment analysis for Ukrainian and Russian news. In their work, 2018 Russian and 2133 Ukrainian news texts were selected for manual annotation using positive, negative, and neutral labels. The study was focused on the reader's perception of the document as a whole, with the question: What sentiments does this text evoke? They also analyzed the relationship between named entities and contexts in the news. In general, research on sentiment analysis for Ukrainian primarily focused on lexicon-based and rule-based approaches, as in [3] and [4] for product reviews or in [5] for political speeches. Babenko and Dyomkin [6] examined applications of machine learning and deep learning algorithms to detect the sentiment of user reviews.

The most recent work of Hamborg and others [7] uses an approach similar to ours, applying BERT's natural language understanding capabilities to political news corpus, but with less detailed corpus annotation rules and for predefined English news fragments.

2. Dataset

2.1. Task of Targeted Sentiment in the News

The goal of sentiment analysis is to computationally identify opinion, emotion, and attitude in the text. It could be done on the document level, on the sentence level, or towards a given target. Aspect-based sentiment analysis relates to the task of extracting fine-grained information by identifying the polarity towards different aspects of an entity in the same unit of text and recognizing the polarity associated with each aspect separately. [8]

The research of algorithms for targeted or aspect-based sentiment in the user review domain was made possible by the availability of datasets, as SemEval 2014 Task 4 [9] and Sentihood [8]. Research on targeted sentiment model architecture includes [10], [11], [12], [13]. Domain-specific BERT language model fine-tuning allowed to achieve state-of-the-art results on the SemEval 2014 Task 4 dataset, with accuracies of 79.19% and 87.14% on the particular subtasks, according to [14]. Still, for the news domain, the evaluation is problematic due to the lack of corpora. Approaches adopted in the analysis of Twitter may not deal well with the specifics of news texts, such as avoidance of emotional words.

In some cases, authors take into account the division of statements into objective and subjective [15], even equating objectivity to a neutral class [16]. This division provides an opportunity to adopt algorithms and lexicons designed to assess the emotionality of the text for subjective

expressions. For example, opinion lexicons were applied to evaluate the set of quotes extracted from news articles [17]. However, the author’s opinion in the news seldom is expressed directly in an emotional way, so the application of sentiment lexicons for opinion mining in the news is limited to special cases. Given the different political views of the readers, possibly different background knowledge of the discussed situation, and the widespread fake news, the task of clearly articulating the concepts of objectivity and subjectivity in the annotation instructions and applying them to targeted sentiment is confusing.

Saif M. Mohammad in [18] gave a comprehensive description of challenges in sentiment annotation and proposed two annotation schemes, suitable for political texts, though not specific to targeted sentiment task: a simple questionnaire with more precise annotation directions and some additional label categories; and a semantic-role based questionnaire.

We decided to focus on an approach that involves a set of questions about each of the pre-marked entities in the text of the news article. These questions concern various aspects of sentiment, and answers can be normalized unambiguously into labels positive, negative, neutral, and uncertain, one per entity.

2.2. Annotation Guidelines

It is typical to think “is it good or bad news?” about a news article. The natural extension of this question for targeted sentiment is the question “is it good or bad news for the entity?” However, it leads to highly subjective annotations, causing problems when one tries to summarize annotators’ opinion.

We formulated the basic approach as “how a statement in a text can affect a public opinion about an entity”. Instead of using the terms “objective/subjective” in annotation guidelines, we asked annotators separately about factual information and emotional vocabulary exposed in a text.

Given the above, we adopted key approaches proposed as “A Simple Sentiment Questionnaire” in [18] to mark emotional opinion about the entity:

1. What kind of language is the speaker using? Is it positive language (expressions of support, admiration, positive attitude) or negative language (for example, expressions of criticism, judgment, negative attitude, questioning validity/competence)?
2. Agreeing or disagreeing with the speaker’s views should not have a bearing on your response. You are to assess the language being used (not the views).

We encouraged the annotators to argue their position looking for specific emotional phrases but decided not to mark them directly because this would significantly increase the cost and the load on the annotators.

For the factual part, we proposed to answer simple questions about the entity like “did the target act responsibly?” or “has the target received public approval?” to mark the facts with positive or negative polarity. To help annotators find the factual statements, we grouped such questions into the table with two columns for positive and negative polarity. Rules for emotional phrases were also organized as a table with questions. We insisted on ignoring the facts or emotions that do not fall under the definition by the tables.

Testing this approach, we found several things that cause difficulties for annotators:

1. Sarcastic or ironic statements are usually hard to evaluate, though in some cases, they are used to express a negative opinion about an entity. To cope with this, we added the ambiguity mark to use with such expressions or when there are both negative and positive assessments or facts. The sentiment was considered neutral if there were no significant positive or negative utterances, as was proposed in [18].
2. In many cases, it was problematic to decide which instance is the target of sentiment. To distinguish between opinion holders and opinion targets, we proposed to check whether an entity is a source of information in this piece of text. Given that the source of information seldom is a sentiment target, that helped clarify the algorithm we suggested to annotators for evaluating entities, cutting off neutral mentions in the first place.
3. Without strictly limiting the fragments for consideration to one sentence, we still asked to evaluate the immediate context of the entity, and not more than 5 sentences at a time.

Thereby, our annotation instruction included three questions:

1. Is the target a speaker (source of information)? Answer options: Yes, No.
2. Does the vocabulary used in the text indicate an emotional assessment? Answer options: Neutral, Positive, Negative, Uncertain.
3. Are there facts in this piece of the text that the entity could consider a success (positive) or a failure (negative)? Answer options: Neutral, Positive, Negative, Uncertain.

This instruction design helped to minimize annotator subjectivity for Ukrainian, Russian, English, and other European languages during the annotation process.

2.3. Document Selection and Annotation

To prepare a targeted sentiment corpus, it was necessary to perform two key tasks:

1. select documents on different topics that with high probability contain non-neutral entities;
2. select the entities in texts that should be annotated.

We used document-level or paragraph-level sentiment models for the first task, assuming that non-neutral entities will be found in the non-neutral context. For the news domain, most of the texts are neutral, and only a small part of the articles is positive, but for the topics like wellness and technology, the diversity is greater. According to this, choosing different news topics helped to balance the corpus.

The task of entity selection was problematic because there was no NER model for Ukrainian with appropriate quality and no corpus big enough to use for model training. The ner-uk corpus [19] was not sufficiently large and was not annotated primarily on news articles. The rule-based NER detection algorithm used by the company at that moment also did not meet our needs. This led to a decision to annotate NER corpora for Russian and Ukrainian, using the documents chosen for the sentiment annotations. For the targeted sentiment, the entities labelled as persons (PERSON), organizations (ORG), products (PRODUCT), and other (MISC) were selected, with MISC label corresponding to works of art, book and movie titles, and events.

Geopolitical entities and locations (LOC) were not used for sentiment annotation but were included at the stage of NER model training.

We used pybossa [20] as an annotation tool. Our annotators were mostly freelance translators and copywriters. For the Ukrainian and Russian NER tasks, we invited five annotators, one of them a supervisor who made a final decision. We reduced this number to three annotators for the sentiment project. This task required more concentration and deep text comprehension from the readers and proved difficult for humans. To track the level of disagreement between annotators on the sentiment project, we monitored the number of choices per entity for each of the three questions separately. The disagreement numbers indicated the necessity of additional work of the super-annotator, who was resolving the ambiguous cases. The labels infosource, fact, and emo were used internally in the process of annotation. Disagreement values (mean number of choices per entity), for Ukrainian and Russian correspondingly, were 1.08 and 1.1 on infosource, 1.14 and 1.12 on fact, 1.1 and 1.09 for emo. We started with 1000 Ukrainian and 1000 Russian documents but excluded some articles from the final corpus. In addition, 150 English documents were annotated for experimentation. Examples of annotation are shown in Table 1.

Table 1

Examples of annotated texts.

Text	Non-neutral or source-of-information entities
Hesham Mansour, Egyptian actor and writer for television shows caused an uproar by posting anti-Semitic statements on Twitter.	Hesham Mansour (infosource_false, fact_negative, emo_neutral)
"We feel very blessed with Curt and his passion for animals and people with disabilities," Leah Wood said.	Curt (infosource_false, fact_neutral, emo_positive) Leah Wood (infosource_true, fact_neutral, emo_neutral)
У спільному розслідуванні The Insider і Bellingcat вдалося встановити особу ключового фігуранта, розшукуваного Спільною слідчою групою у справі про збитий малайзійський Boeing МН17, – це генерал-полковник Микола Федорович Ткачов, головний інспектор Центрального військового округу Росії.	The Insider (infosource_false, fact_positive, emo_neutral) Bellingcat (infosource_false, fact_positive, emo_neutral) Микола Федорович Ткачов (infosource_false, fact_negative, emo_neutral)
"Для нас дуже важливо, що цей проект, який має назву "Шляхи дружби", матиме таку перлину – концерт "Класика заради миру", миру, якого так прагне весь наш народ", – сказав Євген Нищук.	концерт "Класика заради миру" (infosource_false, fact_neutral, emo_positive) Євген Нищук (infosource_true, fact_neutral, emo_neutral)
Державне бюро розслідування: новий ефективний орган чи відстійник старих дискредитованих кадрів.	Державне бюро розслідування (infosource_false, fact_neutral, emo_uncertain)

3. Model

3.1. Preprocessing Data

The first task in preprocessing turned out to be merging the two types of sentiment – emotional and factual – into one. The distinction became hard for the trained models because of the insubstantial number of samples with emotional assessment. For annotators, the division of questions into two groups helped to prescribe the instructions more clearly, but lexical markers of two types of sentiment are often similar in the news, and an entity usually has the same tonality (neutral, positive, negative) on both types of sentiment. For a clear separation of features, annotators have to assess the subtle nuances in the text, which is a difficult task and leads to the additional need of super-annotator work. In the few cases where emotional and factual sentiment did not coincide, we marked the fragments with an "uncertain" label. Ambiguous fragments were removed from the training corpus. One generic sentiment combining the two types is sufficient for completing the task, and an underperforming model is a bigger risk than

Table 2

Annotated entities by sentiment polarity.

Sentiment	Count
Neutral	45296
Positive	2884
Negative	6030
Total	54210

an inability to discern between two types of sentiment.

The annotated dataset consisted of 1958 documents, 987 of them in Ukrainian and 979 in Russian. Each document has, on average, 27.7 named entities (targets for our sentiment analysis). Table 2 shows detailed distribution of data points.

Of the 54210 targets, we used 43775 for the training set, 4864 for the validation set, and 5571 for the test set.

Unlike in document-level sentiment analysis, these targets do not have their own text span that would be unique, clearly delineated, and independent from others. Usually, target sentiment is determined by words and phrases around it, but there is no rule to decide which words exactly should be used as features for predicting the sentiment. What is certain is that we have to pick a “context” – several words or sentences – around the entity and use this context as a clue to the sentiment of that entity. Any such context selection is bound to be based on some rule of thumb. Some of such simple rules might be:

1. Using a whole sentence of which entity is part;
2. Using a window of several sentences around the sentence with the entity;
3. Using a window of multiple tokens around the entity without regard for the sentence boundaries.

Other, more complex strategies are possible, but we tried primarily these three. Of these, all have their problems:

1. Using the sentence as a unit of analysis makes it document-level sentiment analysis, with the same sentiment for all entities contained in a sentence. Any long-range contextual dependencies from neighboring sentences are lost.
2. The second approach is better but suffers from large variability in sentence size. Some contexts might become much longer than others and some much shorter so that in some cases, we have too much irrelevant information in the context, and in others, too little relevant information.
3. The token-based window around the entity deals with the problem of size variance (except for entities at the start of the document and at the end) but leads to sequences with arbitrary beginnings and ends that can leave important words just outside the window.

Despite the disadvantages just mentioned, the third preprocessing strategy performed better than others on the validation set, and we decided to go with it while understanding its limitations.

Table 3
Baseline model results.

Token window size	Model architecture	Embedding size	F1 macro	F1, positive class	F1, negative class
25	Emb + FF	64	0.40	0.12	0.18
25	Emb + FF	128	0.41	0.15	0.21
25	Emb + FF	256	0.42	0.16	0.21
20	Emb + FF	64	0.41	0.14	0.18
20	Emb + FF	128	0.40	0.12	0.2
20	Emb + FF	256	0.42	0.16	0.21

For tokenization, we used a “blank” spacy model for the Ukrainian language. URLs, emails, numbers, whitespace symbols were removed or replaced by special symbols. If there were not enough tokens to the left or to the right of the entity, the window was shortened accordingly.

A particular problem is how to deal with the entity itself – our sentiment target. Any classification approach involving word embeddings would use the target embedding to predict the sentiment of that same target. Moreover, our targets are named entities and do not have the same meanings as usual words, so trying to use their embeddings in any way could lead to noise in the model. That is why we decided to replace all named entities with placeholders according to the entity type: for example, ORG entities were replaced with the word “organization”, PER – with “person”, and so on. All entities in the window, not just the target entity, were processed in this way.

The optimal size of the token window was determined by evaluating the trained model on the validation set. We found that about 20-25 tokens to the left and 20-25 tokens to the right work the best for our data.

3.2. Evaluation Metrics

We use standard classification metrics: precision, recall, and F1. However, F1 on all examples is not informative when classes are not balanced (as in our case). Good performance on neutral fragments can improve F1 even if positive and negative fragments are not identified correctly most of the time. That is why as a primary metric, we use macro F1 – a simple average over F1 scores for each class. We also report F1 for positive and negative classes in the following.

3.3. Baseline

When text is segmented into fragments around named entities and each fragment has a manually annotated label attached, this is a typical classification task. For a baseline, we have chosen a simple feedforward neural network (FF) with one embedding layer (Emb) and one linear layer (see Table 3).

3.4. Using Transformers

When BERT was introduced [21], it promised to improve performance by using a massive pre-trained model based on self-attention layers, which needs to be only fine-tuned – not trained from scratch – to show great results on a wide variety of tasks. That is why BERT (in its multilingual versions) was a natural place to look for improvements.

In BERT, inputs are represented using WordPiece tokenization that segments text into pieces smaller than words. This approach makes it possible to work with large datasets without using a huge vocabulary or embedding each word separately. Because of this, a multilingual BERT model can deal with dozens of languages at the same time. We included both Ukrainian and Russian texts when pre-training the model, so it is a perfect starting model to fine-tune on a bilingual sentiment dataset.

Depending on the model, WordPiece returns a different number of tokens from the text, and this number is larger than the number of words in the text. For BERT, inputs must be of the same length, so it is preferable to set the hyperparameter `max_seq_len` so that most texts in the sample are not truncated and used in full (smaller texts are padded during encoding). Depending on the size of the token window (see above), `max_seq_len` of 128 or 144 works the best for our case.

For fine-tuning, we used an open source transformers library. It allows using BERT for classification tasks by adding a feed-forward neural network with a softmax layer after the pre-trained BERT layers, and then training this network for a few epochs (usually 2-4) on the supervised data,

Apart from the multilingual BERT model, we tried RuBERT pre-trained by DeepPavlov on Russian texts [22]. RuBERT is based on multilingual BERT, so it retains tokens and weights from Ukrainian texts used in pre-training that older model. At the same time, it could improve performance on the Russian fragments. Also, a significant share of roots, prefixes, and suffixes are common to both Russian and Ukrainian, which means that subword embeddings in RuBERT can, in many cases, correctly reflect Ukrainian language use.

4. Results

Comparison of both models – BERT multilingual and RuBERT – are presented in the table below. All models were trained for 3 epochs, and the best checkpoint (based on validation data) was used on testing data, results on which are presented here (Table 4).

We can see that both models perform significantly better than the baseline feedforward network and that differences between them are not significant. RuBERT showed the best macro F1 in the model with `max_seq_len` 128 and the token window of size 25. BERT multilingual is very close when used with `max_seq_len` 144. The difference in the best hyperparameter is probably due to different subword tokenization in RuBERT, which leads to a smaller number of tokens in the input.

We also see that F1 for the positive class is usually worse than for the negative class. This is to be expected for the class with the fewest examples for training. The situation could be improved by a better selection of source data for annotators that would contain more positive mentions of named entities.

Table 4
BERT results.

Pretrained model	Token window size	Max seq length	F1 macro	F1, positive class	F1, negative class
BERT	20	128	0.63	0.47	0.51
BERT	20	144	0.66	0.50	0.56
BERT	25	128	0.6	0.4	0.49
BERT	25	144	0.66	0.54	0.54
RuBERT	20	128	0.65	0.48	0.55
RuBERT	20	144	0.63	0.45	0.53
RuBERT	25	128	0.67	0.53	0.55
RuBERT	25	144	0.66	0.5	0.56

Analyzing the work of the model on the broad range of documents, we can state that the model does not confuse positive and negative examples, which is a priority for our customers. The errors are often in those cases that are also hard for humans. Increasing the number and variety of non-neutral fragments of text would improve the quality for more subtle cases.

5. Future Work

We trained our model for use in production with limited resources, which is why we did not consider models larger than BERT (such as RoBERTa, which also exists in a version trained on Ukrainian texts). For production systems with even more limited resources, a possible solution is using DistilBERT [23] – a smaller and faster model based on BERT. In our experiments, DistilBERT consistently shows macro F1 2-3 percentage points below those of BERT. It is up to the customer to decide whether this is a big enough difference, considering efficiency gains from using the smaller model.

The alternative that would use the same amount of resources but possibly show performance gains is training “own” version of BERT on both Russian and Ukrainian news articles. RuBERT shows that additional pre-training, not just fine-tuning, can have some impact on performance metrics. We are considering going in this direction, despite the obvious problems with resources that it entails.

Currently, the dataset and the models trained on it generalize well to news articles, but not always to other types of text on the internet. One omission is social media. In recent years Facebook and similar platforms have become no less important channels for news content than traditional media, but texts posted there are different in style and in vocabulary used. Specifically for Ukrainian, surzhyk (the vocabulary mix of different languages), dialects, grammatical and spelling errors can be used both accidentally and intentionally to create a certain impression. Also, social media texts often contain allusions and sarcasm, sentiment target is defined less clearly, which is usually hard for annotators. For this reason, we see the necessity to review the methods of document selection for the corpus and improve and test the annotation guidelines in this environment and prepare the separate corpus for targeted sentiment in social media.

6. Conclusion

Targeted sentiment in the context of news articles is a challenging and poorly researched problem in NLP. When we started working on this task, there were no suitable datasets and models that we could use. We developed annotation guidelines that allowed our annotators to create a large dataset in Ukrainian and Russian with each entity marked with neutral, positive, or negative sentiment. We could not exploit the annotations to the full extent because all models had difficulties distinguishing between “fact” and “emotion” in sentiment, given the comparatively small number of negative and positive examples in the corpus.

By using pre-trained BERT models, fine-tuned on this dataset, we achieved good results (macro F1 score of 0.67) and were able to introduce one of these models in the production environment to the satisfaction of our customers.

References

- [1] V. Bobichev, O. Kanishcheva, O. Cherednichenko, Sentiment analysis in the ukrainian and russian news, in: 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), IEEE, 2017, pp. 1050–1055.
- [2] O. Kanishcheva, V. Bobicev, Good news vs. bad news: What are they talking about?, in: RANLP, 2017, pp. 325–333.
- [3] M. Lobur, A. Romaniuk, M. Romanyshyn, Defining an approach for deep sentiment analysis of reviews in ukrainian (2012).
- [4] M. Romanyshyn, The algorithm of deep sentiment analysis of ukrainian reviews, in: Theoretical and Applied Aspects of Cybernetics. Proceedings of the 2nd International Scientific Conference of Students and Young Scientists—Kyiv: Bukrek, 2012.—204 p. ISBN 978-966-399-447-5, 2012, p. 235.
- [5] M. Dilai, Y. Onukevych, I. Dilay, Sentiment analysis of the us and ukrainian presidential speeches, Computational linguistics and intelligent systems (2), 2018 (2018) 60–70.
- [6] D. Babenko, V. Dyomkin, Determining sentiment and important properties of ukrainian language user reviews (2019).
- [7] F. Hamborg, K. Donnay, B. Gipp, Towards target-dependent sentiment classification in news articles, arXiv preprint arXiv:2105.09660 (2021).
- [8] M. Saeidi, G. Bouchard, M. Liakata, S. Riedel, Sentihood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 1546–1556.
- [9] M. Pontiki, D. Galanis, H. Papageorgiou, I. Androutsopoulos, S. Manandhar, M. Al-Smadi, M. Al-Ayyoub, Y. Zhao, B. Qin, O. De Clercq, et al., Semeval-2016 task 5: Aspect based sentiment analysis, in: International workshop on semantic evaluation, 2016, pp. 19–30.
- [10] M. Zhang, Y. Zhang, D.-T. Vo, Gated neural networks for targeted sentiment analysis, in: Thirtieth AAAI conference on artificial intelligence, 2016.
- [11] D.-T. Vo, Y. Zhang, Target-dependent twitter sentiment classification with rich automatic features, in: Twenty-fourth international joint conference on artificial intelligence, 2015.

- [12] J. Liu, Y. Zhang, Attention modeling for targeted sentiment, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, 2017, pp. 572–577.
- [13] C. Sun, L. Huang, X. Qiu, Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence, arXiv preprint arXiv:1903.09588 (2019).
- [14] A. Rietzler, S. Stabinger, P. Opitz, S. Engl, Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification, arXiv preprint arXiv:1908.11860 (2019).
- [15] M. Abdul-Mageed, M. Diab, Subjectivity and sentiment annotation of modern standard arabic newswire, in: Proceedings of the 5th linguistic annotation workshop, 2011, pp. 110–118.
- [16] A. Balahur, R. Steinberger, Rethinking sentiment analysis in the news: from theory to practice and back, Proceeding of WOMSA 9 (2009).
- [17] A. Balahur, R. Steinberger, M. Kabadjov, V. Zavarella, E. Van Der Goot, M. Halkia, B. Pouliquen, J. Belyaeva, Sentiment analysis in the news, arXiv preprint arXiv:1309.6202 (2013).
- [18] S. Mohammad, A practical guide to sentiment annotation: Challenges and solutions, in: Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis, 2016, pp. 174–179.
- [19] lang-uk ner corpus, 2019. URL: <https://lang.org.ua/en/corpora/>.
- [20] Scifabric, Pybossa, 2020. URL: <https://pybossa.com/>.
- [21] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
- [22] Y. Kuratov, M. Arkhipov, Adaptation of deep bidirectional multilingual transformers for russian language, arXiv preprint arXiv:1905.07213 (2019).
- [23] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, arXiv preprint arXiv:1910.01108 (2019).