Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 1613–1621
Marseille, 20-25 June 2022

Resources and Experiments on Sentiment Classification for Georgian

Nicolas Stefanovitch, Jakub Piskorski, Sopho Kharazi
Joint Research Centre, Polish Academy of Sciences, Piksel SRL
nicolas.stefanovitch@ec.europa.eu, jpiskorski@gmail.com, sopho.kharazi@ext.ec.europa.eu

Abstract

This paper presents, to the best of our knowledge, the first ever publicly available annotated dataset for sentiment classification and semantic polarity dictionary for Georgian. We describe the characteristics of these resources and the process of their creation in detail. We also report the results of various experiments on the performance of both lexicon- and machine learning-based models for Georgian sentiment classification. We consider both three- (*positive*, *neutral*, *negative*) and four-tier (*positive*, *neutral*, *negative*, *mixed*) classifications. The machine learning models explored include, logistic regression, support vector machines (SVMs), and transformer-based models. We also explore approaches based on transfer learning and translation (into a well-supported language). The results obtained for Georgian are on a par with state-of-the-art results in sentiment classification for well studied languages when using training data of comparable size.

Keywords: sentiment analysis, low-resourced language, linguistic resources, Georgian language, machine learning

1. Introduction

In this paper we report on creating linguistic resources for the purpose of sentiment detection and classification for the Georgian language, and we compare various knowledge- and machine learning-based models for this task. The main driver behind this work is an extension of the Europe Media Monitor (EMM), a large-scale multilingual news aggregation and analysis engine (JRC, 2018) to the processing of Georgian texts. One important component of the NLP pipeline is sentiment analysis, but there is only little prior work and scarce linguistic resources related to sentiment analysis for Georgian exist. Georgian is significantly under-resourced, making it harder to develop NLP applications for this language. To address this, we developed some resources for performing sentiment analysis in a real-world media monitoring environment.

The main contributions of our work are:

- creating of a sentiment polarity dictionary for Georgian containing circa 2K base and 70K complex entries, ranked on a four-tier scale of expression (*very positive*, *positive*, *negative*, *very negative*);
- creating the first ever publicly released annotated sentiment dataset of ca. 4K text snippets, manually annotated on a four tier scale (*positive*, *neutral*, *negative*, *mixed*) by multiple annotators;
- using these resources to evaluate lexicon- and state-of-the-art machine learning-based sentiment classification approaches.

The machine learning (ML) models explored include, i.a., logistic regression, SVMs, and XLM-Roberta, a multilingual transformer-based model. In particular, we explore different approaches for transformer-based models: direct fine-tuning on the Georgian corpus, fine tuning on the corpus machine translated into English and transfer learning from an already trained model. The experiments consider both three-tier (*positive*, *neutral*, *negative*) and four-tier (*positive*, *neutral*, *negative*, *mixed*) classification. Finally, we perform a detailed study of the Inter-Annotator Agreement (IAA), and evaluate the impact of subsetting the created sentiment-annotated corpus based on IAA values, to explore whether excluding the snippets annotated by the 'outlier' annotators (least in agreement with the others) would improve the performance.

The paper is structured as follows. In Section 2, we present related work. In Section 3, we give an overview of the linguistic resources created, i.e., the semantic polarity lexicon and sentiment-annotated text snippets. Subsequently, in Section 4 we report on the results of the experiments on lexicon- and ML-based approaches to Georgian that exploit the linguistic resources created. Finally, the main findings are summarized in Section 5, which also outlines future work.

2. Related Work

2.1. Georgian Language

Georgian is an isolate language from the Caucasus region, currently spoken by around 3.7 millions people worldwide. It uses its own unicameral script system, it is both agglutinative and inflected with 7 cases, and has some uncommon features such as split-ergativity (different grammatical markers for a given function in the sentence depending on the tense) and polypersonalism.

In Georgian, the verbal forms are derived from a root, to which up to 3 prefixes (preverb, agreement prefix and version vowel), and 4 suffixes can be attached. The preverb plays the same role as the verbal particle in English, potentially totally changing the meaning of the verb. Verbal roots are usually short, they can appear as a substring of other words, and the root can be slightly modified when combined with suffixes. The number of derived forms for a given verbal root can range from several dozen to a few hundred.

Relatively little work has been reported on linguistic resources and NLP tools for Georgian. For instance, (Jassem et al., 2017) reports on the development of basic text processing resources like tokenizers and sentence splitters for Georgian. Work on morphology and a morphologically-tagged corpus for Georgian was reported in (Kardava et al., 2019) and (Nino Doborjginidze, 2016) respectively. (Kapanadze, 2019) and (Kapanadze et al., 2021) report on creating a CFG-based syntactic parser for Georgian. Experiments on text classification of medical reports in Georgian are presented in (Corchado et al., 2016).

2.2. Sentiment Analysis

One of the most basic tasks in sentiment analysis is classifying the polarity of a given text: whether the language used therein is positive, negative or neutral. This can be done at the document, sentence, or feature/aspect level. Both knowledge- and ML-based approaches to sentiment classification have been reported (Liu and Zhang, 2012; Poria et al., 2020; Sudhir and Suresh, 2021). The classic knowledge-based approaches to sentiment classification exploit polarity lexica, i.e. dictionaries of words and/or phrases labelled with semantic orientation (positive or negative), which are used to calculate the overall sentiment of a given text (Taboada et al., 2011; Jurek et al., 2015). The most recent approaches to sentiment classification exploit various ML-based paradigms, ranging from SVMs (Moraes et al., 2013) to deep learning approaches (Zhang et al., 2018). Work on classification of sentiment in short texts, which is the focus of our attention, has been reported in (Kumar et al., 2018; Wang et al., 2018; Wan, 2019).

As previously mentioned, linguistic resources are scarce for the Georgian language, especially for the specific task of sentiment analysis. The only related resource we found to compare against is a short polarity dictionary for Georgian, which is part of a larger project on acquisition of sentiment lexica for many languages (Chen and Skiena, 2014).

While there is no preexisting corpus we can use to train models, multilingual transformer models allow to use transfer learning for Georgian after fine-tuning on training data for sentiment analysis in another language (Barbieri et al., 2021). Another approach used to deal with low-resourced languages is to translate the text into a language for which a good classification model exists (Tebbifakhr et al., 2020). However, to be properly evaluated, both these approaches require the existence of test data in Georgian.

In this context, we were led to create our own linguistic resources in order to develop a good quality sentiment classifier and sentiment classifier evaluation. These resources comprised a sentiment polarity dictionary and a dataset consisting of short text snippets annotated for sentiment.

3. Linguistic Resources

3.1. Sentiment Polarity Dictionary

We have created a sentiment polarity dictionary for Georgian based on the sentiment dictionary for English used in the Europe Media Monitor tool (JRC, 2018), a large-scale multi-lingual news aggregation and analysis engine. Our dictionary has circa 2000 entries, annotated using a four-tier scale of expression: *very positive*, *positive*, *negative*, *very negative*.

For each of the tiers, a Georgian native speaker was asked to translate polarity words from English to Georgian. Words that proved impossible to translate were dropped; and where two English words from two different tiers translated into the same Georgian word, only the word with the strongest polarity was kept.

Because of the complex morphology of Georgian (Ducassé, 2021), it is not possible to directly use the verbal root to recover the different forms pertaining to the same verb. Moreover, the rich morphological productivity of verbal roots makes it extremely time-consuming for an expert to manually list all the different variations. Therefore, we decided to use a template system for verbs in order to generate the derived forms.

3.1.1. Generative template system

We automatically derive all possible morphemes of a verb, based on its root and up to two additional parameters: (a) a list of potential preverbs, and (b) a dependent noun. A dependent noun is used for compound verbs, verbs that require the use of a noun, for instance, the English verb ‘*to water proof*’. The annotation scheme for encoding the verbs with all this information is outlined below:

```
VB: root
VB: (preverb list)+root
VB: dependant noun+(preverb list)+root
```

Derived forms are matched using a combination of wildcard expansion for suffixes and explicit generation for combination of prefixes. Prefix generation includes: preverb, agreement prefix and version vowel. The last two take their values from closed lists, without changing the meaning of the verb, and are therefore not part of the parameters of the patterns. For instance, the verb ‘*to love*’ and all its variants are encoded as `VB:`(შე)+ყვარ. Not all the forms generated exist, however, this does not pose a problem since the non-existent forms would not be found in texts and therefore would not impact the quality of sentiment analysis.

Handling derived forms of nouns or adjectives was straightforward: the analyst had to provide the stem of a word, and a wildcard was assumed at the end of the word, e.g., the noun ‘*love*’ is encoded as ‘სიყვარულ’.

The template system also makes it possible to specify exceptions: to include or exclude specific forms whenever deemed necessary. Furthermore, to deal with negation on a basic level, where a polarity word immediately followed a negative marker, we automatically expanded the sentiment dictionary by linking 4 negative connectors (არა, არ, ვერა and ვერ) to all polarity words and inverting their polarity.

The statistics for the polarity lexicon resulting from applying all the steps above are provided in Figure 1.

	Very Pos.	Pos.	Neg.	Very Neg.	Total
RAW	84	721	831	350	1986
EXPANDED	342	4220	6989	2572	14123
FINAL	10630	32176	23869	3940	70615

Figure 1: The statistics of the sentiment polarity dictionary for the different tiers of polarity: RAW raw corresponds to the number of templates, EXPANDED row depicts the number of morphemes after dictionary expansion, and FINAL raw shows the final size of the dictionary after expansion of word forms with negative connectors.

3.2. Georgian Sentiment Snippets

In order to facilitate training and evaluating ML-based sentiment classifiers for Georgian, we created a new resource for this purpose, which contains 4223 text snippets, annotated using 4 categories: *positive*, *neutral*, *negative* and *mixed*, with the latter used to annotate snippets containing both positive and negative sentiment triggers. Twelve native-speaker annotators were involved in this task. The following subsections describe the entire process in more detail.

3.2.1. Data sampling

To create input material for annotations, we extracted full sentences in Georgian from the archive of Georgian news articles gathered by EMM[1]. News articles were randomly sampled over a period of 5 years: for each month, 100 news articles were randomly selected and sentences were extracted. If a sentence contained quotations, the sentence was further split around the quotation boundaries. From this pool of sentences, we randomly sampled the corpus snippets. Based on empirical observations, revealing that the resulting sentences were highly unbalanced in terms of potential sentiment category, the sampling process was modified and biased to obtain a more balanced dataset. The entire set of sampled snippets is divided into three subsets:

- I: snippets sampled without any additional constraints (3702 snippets);
- II: snippets sampled with a requirement to contain the most frequent names, with the assumption that they constitute polarising entities, and thus, increase the probability of non-neutral sentiment (although preliminary results did not support this hypothesis) (126 snippets);
- III: snippets sampled with a requirement to contain at least one positive word from the sentiment dictionary described in 3.1 (394 snippets).

3.2.2. Data annotation

The annotators were asked to annotate each text snippet using either one of the four sentiment labels or an *Ignore* tag where the text is corrupt and should be discarded. Given that most of the annotators were not experienced in terms of prior annotation work, and to take extra care to avoid political bias altering the sentiment evaluation, the annotators were asked to tag the objective sentiment (as expressed by the language used) separately from the subjective sentiment that the news evokes in the reader. The annotators underwent two rounds of training before being given the actual data. On average, they gave the same label to the objective and subjective sentiment tags in 86.3% of the cases.

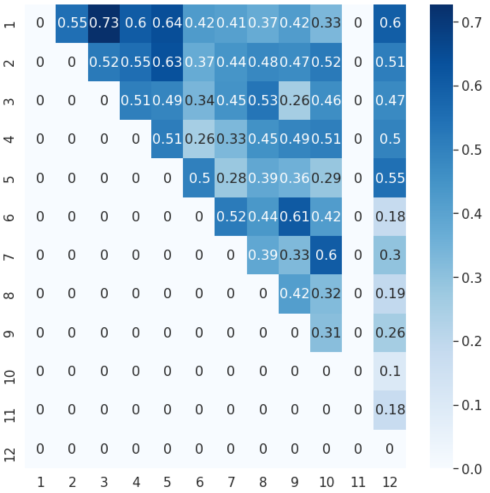

Figure 2: IAA: Cohen's κ for all pairs of annotators with at least 15 annotations in common for the three-tier sentiment classification. Annotators are ordered by their average agreement from 1 to 11; the expert annotator is the 12th annotator

Each snippet in sets I and II (see Section 3.2.1) was first labelled by two annotators. Before resolving conflicts, the annotation agreed for only circa 50% of the snippets. The snippets from set III were annotated by a single experienced annotator. The same annotator resolved the conflicts for the non-agreeing annotations for sets I and II.

3.2.3. Agreement

The snippets were allocated to annotators in such a way that all pairs of annotators could be compared. To measure the Inter Annotator Agreement (IAA) for pairs of annotators, we used Cohen's kappa κ. We computed this value for all pairs of annotators with at least 48 annotations in common for the three-tier analysis; the results are provided in Figure 2. On average, there were 58 annotations in common. Eleven annotators participated in the initial annotation phase (`All`). An additional expert annotator (`Expert`) annotated all the snippets with disagreement and also provided annotations for additional snippets. The expert has a high IAA with all the top 5 annotators (`Top5`), who also all have a high IAA with each other.

To measure the IAA for the entire dataset, Krippendorff's α was used, as it is a better indicator of overall agreement than Cohen's κ (Zapf et al., 2016). For the three-tier sentiment classification, the α value was 0.461 for `All`, 0.571 for `Top5`, 0.622 for

[1] `http://newsbrief.eu`

`Top5+Expert`, and 0.543 for `All+Expert`. According to (Hayes and Krippendorff, 2007) a value above 0.667 is recommended for acceptable agreement. Despite being slightly under that limit, the α value is on a par with the best values achieved by (Mozetič et al., 2016), who created a sentiment analysis dataset for different languages.

3.2.4. Data perturbation

Given that the annotated snippets often contain named entities, various perturbation techniques were applied to: (a) introduce more variation into the final version of the corpus, and (b) reduce potential political bias. These perturbations include:

- changing all numerical expressions and some temporal expressions represented as numbers, by adding/subtracting some random number;
- replacing the most frequent person names detected in the snippets, by randomly selecting a replacement from a pool of names computed over the whole data gathered before sampling snippets preserving the inflection of the names to ensure grammatical correctness;
- randomly replacing a limited set of frequent country names with alternative country names, while preserving the case markers;
- manually perturbing some text snippets where none of the above techniques could be applied, e.g., changing adjectives, changing named-entities, replacing words with synonyms, etc.

Overall, 56.7% of the text snippets were modified using the perturbations described above, with at least 75% of these perturbations resulting from changing at least one named entity in the text snippet. As a consequence of the perturbations, a large fraction of the resulting text snippets do not fully correspond to the description of certain real-world events.

3.2.5. Quantitative description

Figure 3 presents the statistics on the composition of the final annotated dataset, which will be referred to as Georgian Sentiment Snippets (GSS). To make the dataset as balanced as possible, a variety of sentence sampling approaches were used (see Section 3.2.1). Subset II did not provide a significantly different distribution of labels to subset I, with *neutral* again being the predominant class, hence the recourse to subset III. Despite all these efforts, the dataset remains unbalanced: the *neutral* class represents 41% of the snippets, and the *negative* class is twice as populated as the *positive* class. Figure 4 provides the text length-distribution histogram for the annotated dataset. The average length of the text snippets is 114 characters.

Perturbation of named entities has a regularization effect. In Table 1, we report the class distributions obtained by filtering the snippets to contain the top 10 last names (**10-L**) and the top 100 first names (**100-F**), each applied to both the original and perturbed dataset. The resulting distributions are compared with the full label distribution using Jensen-Shannon divergence, taking into account only the three main classes. We can observe that, in the original dataset, there is an overall positive bias towards the top 10 entities as the perturbations increase their proportion of *negative* labels by 5.9% and decrease the proportion of *positive* labels by 1.2%. When considering the top 100 first names that effect is less pronounced, and in both cases the perturbed dataset is closer to the exact label distribution, i.e., the bias linked to these specific entities has been reduced.

Negative	Neutral	Positive	Mixed	Total
1417	1734	765	307	4223
33.5%	41.0%	18.1%	7.2%	100%

Figure 3: The statistics of the GSS corpus.

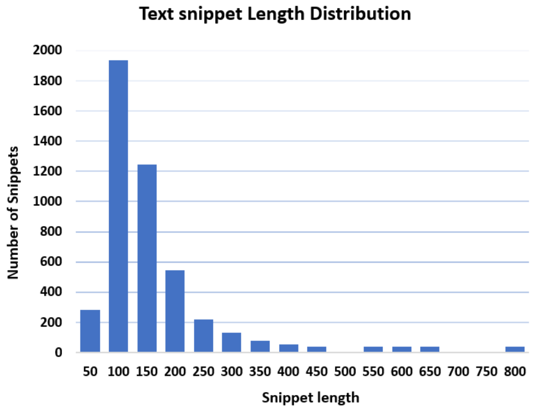

Figure 4: Text snippet length distribution.

4. Experiments

This section presents the results of the evaluation of both lexicon-based (see Section 4.1) and ML-based approaches (see Section 4.2). To measure sentiment classification performance we used $precision$, $recall$, and the $micro$, $macro$ and $weighted$ F_1 metrics.

4.1. Dictionary-based Approaches

Polyglot, a statistically obtained lexicon of 886 positive and 1316 negative sentiment words described in (Chen and Skiena, 2014). We used a simple algorithm to check for the presence of sentiment words in a given text and assign the respective sentiment class, where the presence of both positive and negative words results

Experiment	Neg.	Neut.	Pos.	JS div.
none	36.2%	44.3%	19.5%	0.0
10-L in Pert.	38.9%	42.9%	18.2%	4.1e-4
10-L in Orig.	33.0%	47.6%	19.4%	6.6e-4
100-F in Pert.	36.6%	47.4%	16.0%	1.1e-3
100-F in Orig.	35.2%	48.6%	16.2%	1.3e-3

Table 1: Impact of named entity perturbations on the label distribution of top entities

Table 2: The Evaluation results for dictionary-based approaches.

	Micro average			Macro average			Weighted average		
System	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
Polyglot	49.6	49.6	49.6	47.4	49.3	47.1	49.6	50.4	48.5
GL-2	53.1	53.1	53.1	54.8	54.2	52.8	53.1	55.5	52.9
GL-4	56.1	56.1	56.1	56.2	**59.9**	56.1	58.8	56.1	56.0
GL-4N	56.0	56.0	56.0	56.3	**59.9**	56.1	58.9	56.0	56.0
GL-4-LR	59.9	59.9	59.9	60.4	56.6	57.8	60.2	59.9	59.5
GL-4N-LR	**60.4**	**60.4**	**60.4**	**60.8**	56.7	**57.9**	**60.7**	**60.4**	**59.9**

in a *neutral* score. Given that Polyglot lexica were the only linguistic resources we found for semantic analysis of Georgian, the Polyglot-based approach constitutes the baseline here for evaluation purposes. Polyglot sentiment lexica for Georgian contain some stopwords as well as English words. Nevertheless, for the sake of consistency we used these lexica 'as are'.

Georgian Polarity Dictionary, derived from English sentiment lexicon (see Section 3.1), with 4 polarity labels. We computed the polarity score of a sentence as the weighted average of the words present in the dictionary, where a positive score yields *positive*, a negative score yields *negative* and a null score yields *neutral*. The score of a given word w is 0 if no match can be found for w in the sentiment dictionary, +1 if w matches a positive word in the dictionary, and -1 if w matches a negative word in the dictionary. Using the GSS corpus as test data, we performed several experiments using the expanded dictionary. Words were matched if the entry in the dictionary exactly matched the beginning of a word, this meant that words were matched irrespective of inflection. In the first experiment, we used the scoring outlined above, hence, considering only 2 possible polarities (**GL-2**). In the second experiment, we took into account the intensity of the polarity by doubling the score of a word if it had stronger polarity (**GL-4**). In the third experiment, we did as described previously, and additionally expanded the dictionary with the negative connectors, inverting the polarity of a word where they appeared directly before it (**GL-4N**). Finally, we considered hybrid approaches, i.e., we used the raw counts provided by GL-4N, and expanded this with additional features representing the proportion of words for each polarity, with respect to the total number of words and the total number of matched words. These features were used to train a logistic regression model, resulting in two approaches **GL-4-LR** and **GL-4N-LR**, depending on whether negation was taken into account.

The performance of the dictionary-based approaches is presented in Table 2. The entire GSS dataset was used for evaluation, except in the GL-*-LR approaches, where snippets from subset III were excluded to avoid the bias due to the tonality dictionary being used to select that sample. For these algorithms, requiring training, a 5-fold cross validation was carried out. Among the non-neutral snippets in GSS, 15.7% did not contain any word in the polarity dictionary.

4.2. Machine Learning-based Approaches

We explored various ML paradigms, for both three-tier and four-tier sentiment classification, and used 5-fold cross-validation with an 80:20 split, unless otherwise specified. The ML-based approaches are:

L2-regularized Logistic Regression (L2-LR) with 3-6 character n-grams found in the texts as binary features[2], vector normalization and $c = 20.0$ and $\epsilon = 0.05$ resulting from parameter optimization.

Support Vector Machines (SVM) with 3-6 character n-grams as binary features [3], vector normalization and $c = 0.7$ for three-tier sentiment classification and $c = 1.0$ for four-tier, $\epsilon = 0.05$, and $p = 0.1$ resulting from parameter optimization. We used the version of the SVM algorithm with a linear kernel described in detail in (Crammer and Singer, 2000). We have also translated the entire GSS corpus into English[4] and trained an SVM model using this data, which will be referred to as SVM-EN. LIBLINEAR library[5] was used to carry out experiments with L2-LR and SVMs.

Transformers (TF-*)
For all the experiments using transformers, we used the XLM-T language model (Barbieri et al., 2021), which is based on the model `xlm-roberta-base`. XLM-T's tokenizers include Georgian, and it has undergone further pre-training based on 200M tweets in over 30 languages. We fine-tuned this model for the sentiment classification task using various datasets. We chose the model as it is one of the top performers for multilingual transfer learning. XLM-T was chosen over the larger language model `xlm-roberta-large` because preliminary experiments showed that the latter often demonstrated high variability in performance, while XLM-T did not seem to be subject to that effect. The training data used was either the GSS corpus (see Section 3.2) or the Unified Multilingual Sentiment Analysis Benchmark (UMSAB) dataset, which is used

[2] An n-gram is considered as a feature only if it appears at least 3 times in the training data.

[3] Log-scaled TF-IDF weighting yielded similar results.

[4] We used Google API for all translations

[5] `https://www.csie.ntu.edu.tw/~cjlin/liblinear`

by XML-T authors to fine-tune their language model. UMSAB covers 8 languages and contains about 24K tweets. We pre-processed this corpus to remove all hashtags, usernames and URLs; remaining sentences shorter than 5 characters were then discarded. The texts are of a different nature from the news, however, they are comparable in terms of size.

We carried out the following experiments: training and testing on GSS only (**TF-GG**), training and testing on GSS translated into English (**TF-GG-EN**), training on UMSAB and testing on GSS (**TF-UG**), training on UMSAB and testing on GSS translated into English (**TF-UG-EN**), training and testing on a combination of GSS and UMSAB (**TF-HH-5** and **TF-HH-35** with a respective minimal length of 5 and 35 characters). For reference, we also provide the results obtained when training and testing on UMSAB alone (**TF-UU**).

These models were trained with 3 epochs and a batch size of 32. Where the test and training datasets are the same, 5-fold cross validation is used, otherwise, the full datasets are used respectively for training and testing.

4.2.1. Three-tier classification

The overall performance results for ML-based approaches to the three-tier sentiment classification task, trained and evaluated on the GSS corpus are presented in Table 3. The accuracy of a random guess, based on the distribution of the three classes in the GSS corpus, is 37%. Not surprisingly, the transformer-based approach achieved the best overall results (75.6 and 75.2 weighted and macro F_1 score resp.).

In Figure 4.2 we present the confusion matrices for the TF-GG and SVM models. The errors depict understandable behaviour, with most confusion between positive and neutral, and neutral and negative, for both SVM and transformers.

The translation-based approach performs surprisingly well, with SVM-EN lagging only 1.4 points in macro F_1 score behind its SVM counterpart trained on GSS data. TF-GG-EN actually performs better: up to 1.6 points in macro F_1 score over its pure "Georgian" counterpart. This difference in performance could be due to the significant difference in vocabulary size between English and Georgian in `xlm-roberta-base`. In terms of the number of tokens containing strictly letters of the respective alphabets we get a tokens count of 83017 tokens for English and 3770 tokens for Georgian: a 22 fold difference. Training loss decreased faster for TF-GG-EN and testing loss started to increase around the second epoch, which was not the case for TF-GG, indicating a possible overfitting of TF-GG-EN. This would likely not occur with a comparable set of features for both languages.

The results of the transformer-based approaches that use different datasets for training and for evaluation are presented in Table 4. TF-UU allows us to compare the training of our model with the paper reference, and confirms that they display similar overall performance. The experiment TF-UG is intended to study the performance of transfer learning and its adequacy for tackling under-resourced languages. TF-UG shows that the model performs poorly at transfer learning into Georgian. While performing better than a random guess, its weighted F_1 performance is about 30 points lower than TF-GG, and up to 10 points lower than the simplest lexicon-based approach, GL-2.

TF-UG-EN shows that, when independently training on UMSAB and testing on GSS translated into English, performance is on a par with TF-UU. The L2-LR and SVM-based solutions constitute an interesting alternative to TF-UG-EN, lagging only a point behind in macro F_1 average.

TF-HH combines both datasets for training and testing, where filtering by sentence size yields a 2 point gain, indicating that we could expect performance gain from applying such filtering in other UMSAB-trained models.

For the sake of completeness, we have compared the results of some of the approaches, including GL-2, SVM and TF-GG, on the non-perturbed versus perturbed data, with the results provided in Table 5. We can observe that the results obtained on the perturbed version of the corpus are worse by only 0.4-0.9 points. We attribute this to the regularisation effect of random perturbation of frequent named entities, as described in Section 3.2.5, lowering the performance on the test set.

4.2.2. Four-tier classification

Most work reported on sentiment polarity focuses on three-tier classification. Given that GSS contains snippets labelled *mixed*, which in fact, reflects the presence of such texts in real-world news, we have also carried out experiments in training four-tier sentiment models. The respective performance of L2-LR, SVM, SVM-EN and TF-GG classifiers trained and evaluated on GSS is presented in Table 6. The accuracy of a random guess, based on the distribution of the 4 classes in GSS is 32%. The confusion matrix for TF-GG for four classes is shown in 4.2. The distribution of mis-prediction for the three classes (excluding *mixed*) is fairly similar to that for the corresponding three-tier sentiment classifiers. The *mixed* class is almost never predicted, while snippets labelled *mixed* are predicted as *positive* and *negative* at a much higher rate than *neutral* wrt. their distribution in GSS. This seems to indicate that for the classifier there are no *mixed* cases, and that what gets predicted is the dominant polarity of a sentence.

4.2.3. IAA-based subsets

In Section 3.2.3 we saw that there is high variability in IAA scores between pairs of annotators. To assess whether excluding the snippets annotated by the worst performing annotators would improve the performance of the trained models we trained additional models on the following subsets of snippets: (a) `All` - snippets annotated by anybody except the expert, and (b) `Top5` - snippets annotated by at least one of the top 5 an-

Table 3: Evaluation results for ML-based approaches for three-tier sentiment classification using the GSS corpus for training and evaluation.

	Micro average			Macro average			Weighted average		
System	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
L2-LR	67.1	67.1	67.1	64.8	68.3	66.1	67.1	67.4	66.9
SVM	67.2	67.2	67.2	65.3	68.0	66.3	67.2	67.3	67.0
TF-GG	**75.6**	**75.6**	**75.6**	**76.8**	**75.0**	**75.2**	**76.0**	**75.6**	**75.6**
SVM-EN	66.3	66.3	66.3	63.7	67.3	64.9	66.3	66.6	66.1
TF-GG-EN	77.0	77.0	77.0	76.8	77.1	76.8	77.1	77.0	76.9

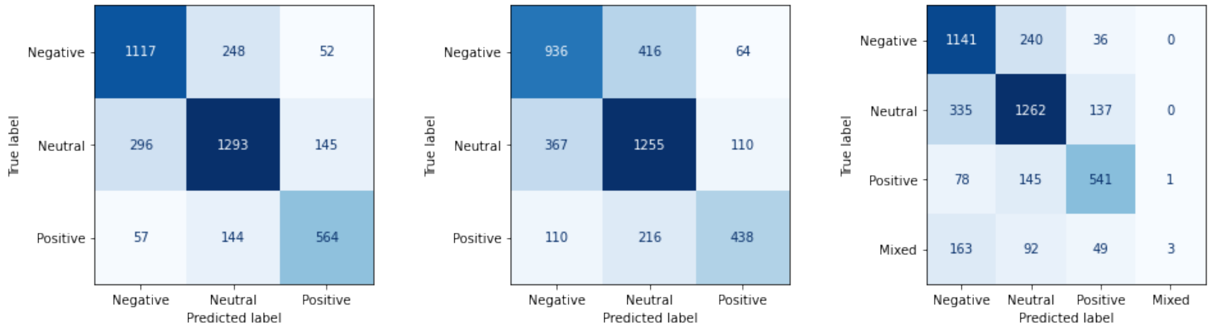

Figure 5: Confusion matrices: (a) TF-GG for three classes (LEFT), (b) SVM for three classes (MIDDLE), and (c) TF-GG for four classes (RIGHT)

Table 4: Evaluation results for ML-based approaches for three-tier sentiment classification using various corpora for training and evaluation.

	Micro average			Macro average			Weighted average				
System	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Train Dataset	Test Dataset
TF-UG	52.0	52.0	52.0	73.0	44.2	40.7	68.0	52.0	43.4	UMSAB	GSS
TF-UG-EN	68.7	68.7	68.7	75.1	65.3	67.5	72.7	68.7	68.1	UMSAB	GSS (EN)
TF-UU	66.8	66.8	66.8	66.9	66.7	66.5	66.9	66.8	66.5	UMSAB	UMSAB
TF-HH-5	66.6	66.6	66.6	66.9	67.0	66.4	66.9	66.6	66.2	UMSAB+GSS	UMSAB+GSS
TF-HH-35	68.1	68.1	68.1	68.4	68.3	67.8	68.4	68.1	67.6	UMSAB+GSS	UMSAB+GSS

Table 5: F_1 scores for some ML-based approaches on non-perturbed versus perturbed data.

	Non-perturbated data			Perturbated data		
System	Micro	Macro	Weighted	Micro	Macro	Weighted
GL-2	53.1	52.8	52.9	52.7	52.3	52.5
SVM	67.6	66.7	67.5	67.2	66.3	67.0
TF-GG	75.9	75.6	76.9	75.6	75.2	75.6

Table 6: Evaluation results for ML-based approaches for four-tier sentiment classification on the GSS corpus.

	Micro average			Macro average			Weighted average		
System	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
L2-LR	61.8	61.8	61.8	49.6	52.4	49.8	61.8	59.7	60.2
SVM	61.0	61.0	61.0	49.4	52.1	49.6	61.0	59.0	59.6
SVM-EN	59.8	59.8	59.8	38.3	**59.8**	38.2	59.8	57.5	58.2
TF-GG	**69.9**	**69.9**	**69.9**	**57.5**	56.7	**55.2**	**66.6**	**69.9**	**67.5**

notators of `All`-annotated snippets. We focus on the top 5 annotators, because pair-wise comparison indicated high agreement between them, as well as with the expert annotator. Performance of the transformer-based models on the three-tier case for all these datasets and their combination with the dataset `Exp` (snippets annotated by the expert annotator), is considered both with and without data perturbation. This is reported in Table 7, along with corresponding Krippendorff's α scores.

Table 7: IAA experiments

		Micro average			Macro average			Weighted average			
Experiment	Support	Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1	α
All+Exp & Pert.	3916	75.6	75.6	75.6	75.8	75.0	75.2	76.0	75.6	75.6	0.543
Top5+Exp & Pert.	3451	76.5	76.5	76.5	76.8	76.4	76.4	76.9	76.5	76.5	0.622
All+Exp	3916	75.9	75.9	75.9	75.8	75.5	75.6	76.1	75.9	75.9	0.543
Top5+Exp	3451	76.9	76.9	76.9	77.3	76.6	76.8	77.1	76.9	76.9	0.622
All	1818	71.2	71.2	71.2	42.5	46.1	43.7	61.7	71.2	65.5	0.461
Top5	1401	67.3	67.3	67.3	52.8	46.7	44.4	66.7	67.3	62.3	0.571

Comparing `Top5` and `All` shows that relatively high agreement in some cases does not automatically imply higher classification performance, as quantity of data is also important in itself. However, when combined with `Exp`, `Top5+Exp` has the highest α, and also the highest performance vis-a-vis other settings. This confirms the idea that selecting snippets from top annotators does indeed improve model performance. `All+Exp`, despite having 13.4% more data than `Top5+Exp`, performs only about 1 point worse. This indicates that a lower IAA is more a sign of wasted resources on annotations than it is a threat to performance.

5. Summary

In this paper, we introduced two linguistic resources for sentiment analysis in Georgian: (a) a semantic polarity lexicon for Georgian, and (b) GSS, a corpus of around 4K text snippets in Georgian labelled with sentiment tags using a four-tier scale (*positive*, *negative*, *neutral* and *mixed*). We studied their properties and also reported on the evaluation of lexicon- and state-of-the-art ML-based approaches for the task of sentiment classification in Georgian. The main findings of the experiments can be summarized as follows:

- Using a purely lexicon-based approach, with the top-ranking setting we achieve a macro F_1 score of 56.1.
- An XLM-Roberta-based model achieved the highest macro F_1 score (75.2) among the ML-based models trained and evaluated solely on the new corpus introduced in this paper, outperforming the classic char-ngram based linear SVM with a macro F_1 score of 66.3.
- XLM-Roberta presented poor transfer-learning capabilities for Georgian, with performance lower than a simple lexicon-based approach. When using a pre-trained model evaluated on GSS translated into English, performance was on a par with the overall performance of the pre-trained classifiers for other languages, but lower than when training and testing on GSS (either directly in Georgian or after translation into English).
- The transformer-based model trained on a version of the Georgian corpus translated into English yielded better results (macro F_1 of translated version). However, given the level of difference in respective vocabulary sizes in XLM-Roberta, this is likely an overfitting effect due to faster convergence.
- The best macro F_1 result obtained for the four-tier classification with XLM-Roberta is quite low (55.2), due to the difficulty of correctly predicting the instances of the *mixed* class that the model tends to assign to the dominant polarity. Given that the *mixed* class is relatively low populated, it is not possible to draw strong conclusions from the the four-tier classification.
- Narrowing down the GSS corpus to snippets annotated by top-agreeing annotators leads to marginal improvement in classification performance.

We believe the linguistic resources for Georgian sentiment classification[6] and the findings of the evaluation of various knowledge- and ML-based models for the task presented in this paper constitute useful material for researchers working on sentiment analysis in Georgian language. We also believe they will be of interest to researchers studying the applications of ML for rare and under-resourced languages.

Transfer learning did not prove satisfactory enough to adequately support Georgian. Where available, translation into a well-supported language proves an effective solution only if the resources in the target language are similar in terms of content to those of the application domain. Where these conditions are not met, the simplest way to obtain high-performance classifiers is to develop ad-hoc resources for the language/task at hand.

In terms of future research into Georgian sentiment classification, we plan to: (a) pre-train transformer-based language models with Georgian data, (b) explore four-tier sentiment classification in greater depth and tackle the hard case of the *mixed* class, (c) elaborate document-level sentiment classification by aggregating sentence-level sentiment scores, and (d) explore solutions for topic-specific sentiment analysis.

6. Acknowledgements

We are greatly indebted to all annotators, who were native speakers of Georgian and without whom the presented work would not have been possible.

[6] The resources are available at: `https://data.jrc.ec.europa.eu/dataset/9f04066a-8cc0-4669-99b4-f1f0627fdbbf`

7. References

Barbieri, F., Espinosa-Anke, L., and Camacho-Collados, J. (2021). A Multilingual Language Model Toolkit for Twitter. In *arXiv preprint arXiv:2104.12250*.

Chen, Y. and Skiena, S. (2014). Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, Baltimore, Maryland. Association for Computational Linguistics.

Corchado, J. M., Khachidze, M., Tsintsadze, M., and Archuadze, M. (2016). Natural language processing based instrument for classification of free text medical records. *BioMed Research International*, 2(1):37–53.

Crammer, K. and Singer, Y. (2000). On the learnability and design of output codes for multiclass problems. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, COLT '00, page 35–46, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ducassé, M. (2021). Kartu-Verbs: A Semantic Web Base of Inflected Georgian Verb Forms to Bypass Georgian Verb Lemmatization Issues. In Zoe Gavriilidou, editor, *XIX EURALEX conference*, 1st volume of XIX EURALEX conference proceedings, Alexandroupolis, Greece, November.

Hayes, A. F. and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication methods and measures*, 1(1):77–89.

Jassem, K., Tsikarishvili, I., Otskheli, T., and Boryczka, U. (2017). On the development of nlp tools for the georgian language. In P. Paroubek Z. Vetulani, editor, *Proceedings of the 8th Language and Technology Conference Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 327–331.

JRC. (2018). Europe media monitor. Technical report, European Commission Joint Research Centre, Ispra.

Jurek, A., Mulvenna, M., and Bi, Y. (2015). Improved lexicon-based sentiment analysis for social media analytics. *Security Informatics*, 4(9):1–13.

Kapanadze, O., Kotzé, G., and Hanneforth, T. (2021). Building resources for georgian treebanking-based nlp. *Lecture Notes in Computer Science*.

Kapanadze, O. (2019). Parsing the less-configurational georgian language with a context-free grammar. In *Proceedings of the 1st International Conference on Language Technologies for All*, pages 342–345, Paris, France. European Language Resources Association (ELRA).

Kardava, I., Gulua, N., Antidze, J., and Toklikishvili, B. (2019). Morphological synthesis and analysis of georgian words. pages 232–235.

Kumar, H. M. K., Harish, B. S., Kumar, S. V. A., and Aradhya, V. N. M. (2018). Classification of sentiments in short-text: An approach using msmtp measure. In *Proceedings of the 2nd International Conference on Machine Learning and Soft Computing*, ICMLSC '18, page 145–150, New York, NY, USA. Association for Computing Machinery.

Liu, B. and Zhang, L., (2012). *A Survey of Opinion Mining and Sentiment Analysis*, pages 415–463. Springer US, Boston, MA.

Moraes, R., Valiati, J. a. F., and GaviãO Neto, W. P. (2013). Document-level sentiment classification: An empirical comparison between svm and ann. *Expert Syst. Appl.*, 40(2):621–633, feb.

Mozetič, I., Grčar, M., and Smailović, J. (2016). Multilingual twitter sentiment classification: The role of human annotators. *PloS one*, 11(5):e0155036.

Nino Doborjginidze, I. L. (2016). Corpus of the georgian language. In George Meladze Tinatin Margalitadze, editor, *Proceedings of the 17th EURALEX International Congress*, pages 328–334, Tbilisi, Georgia. Ivane Javakhishvili Tbilisi University Press.

Poria, S., Hazarika, D., Majumder, N., and Mihalcea, R. (2020). Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research. *IEEE Transactions on Affective Computing*.

Sudhir, P. and Suresh, V. D. (2021). Comparative study of various approaches, applications and classifiers for sentiment analysis. *Global Transitions Proceedings*, 2(2):205–211. International Conference on Computing System and its Applications (ICCSA-2021).

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37:267–307, 06.

Tebbifakhr, A., Negri, M., and Turchi, M. (2020). Machine-oriented nmt adaptation for zero-shot nlp tasks: Comparing the usefulness of close and distant languages. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 36–46.

Wan, Y. (2019). Short-text sentiment classification based on graph-lstm. In *2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM)*, pages 35–38.

Wang, J.-H., Liu, T.-W., Luo, X., and Wang, L. (2018). An LSTM approach to short text sentiment classification with word embeddings. In *Proceedings of the 30th Conference on Computational Linguistics and Speech Processing (ROCLING)*, Hsinchu, Taiwan.

Zapf, A., Castell, S., Morawietz, L., and Karch, A. (2016). Measuring inter-rater reliability for nominal data–which coefficients and confidence intervals are appropriate? *BMC medical research methodology*, 16(1):1–10.

Zhang, L., Wang, S., and Liu, B. (2018). Deep learning for sentiment analysis : A survey. *CoRR*, abs/1801.07883.