*Human Language Technologies – The Baltic Perspective*
*A. Utka et al. (Eds.)*


*doi:10.3233/978-1-61499-442-8-83*

# Normalization and Automatized Sentiment Analysis of Contemporary Online Latvian Language

Ginta GARKĀJE[a,1], Evelīna ZILGALVE[a] and Roberts DARĢIS[a]
[a]*Institute of Mathematics and Computer Science, University of Latvia*

**Abstract.** We describe contemporary language transliteration influence on automatized sentiment analysis. We state that the text normalization helps to achieve better results in automatized sentiment analysis and provide results to support the claim. Data used for the experiments are gathered via project *Virtual Aggression Barometer*. We use a normalization tool and an automatized classifier for the internet user comments with aggressive and non-aggressive sentiment.

**Keywords:** Online language, Transliteration, Normalization, Sentiment analysis

## Introduction

Sentiment analysis and normalization are among significant topics in natural language processing. Sentiment analysis is used in a specific sense in the project *Virtual Aggression Barometer*[2]. Typical positive, negative and neutral classification is replaced by looking for aggressive and non-aggressive comments. The research problem of distinguishing between these two categories is interfered by contemporary online language transliteration, lack of syntactic features and overall low quality of the corpus. The data used in this research consist of user comments from the most popular Latvian news portals: *Apollo*, *Tvnet* and *Delfi* comprising over 11 million user comments.

Normalization tool[3] was used to correct the transliteration mistakes. The sentiment analysis was performed using classical Naïve Bayes classifier library [1]. Machine learning experiments with various normalization setups were run to evaluate the significance of the text normalization in sentiment analysis.

[1] Corresponding Author: Ginta Garkāje, aInstitute of Mathematics and Computer Science, University of Latvia; E-mail: ginta@ailab.lv
[2] http://barometrs.korpuss.lv/
[3] Source code available: https://bitbucket.org/Ginta/ruukjiishi

## 1. Sentiment Analysis

An initial study to find the class proportion was carried out to conduct a Naïve Bayes classifier training in distinguishing aggressive comments from the non-aggressive ones. Social science students annotated three thousand comments (randomly chosen from the comment corpus) dividing them into three categories: aggressive, non-aggressive and neutral. Each comment was annotated two times, and if the opinions did not match, the third annotator made the final vote. There were two conclusions made after this procedure. First, the inter-annotator agreement level is rather low – only 78%. Second, only 17% of the three thousand comments were classified as aggressive.

According to the inter-annotator agreement, the overall upper bound of the classifier accuracy was established as 78%. Furthermore, when using Naïve Bayes classifier, it is important not to exclude features (in this research only words were used as features).

The natural distribution of the aggressive and non-aggressive comments raised a question on the best way to gather the training data. There was a risk that the classifier would learn the trivial notion to classify all comments as non-aggressive (the result would be 83% in overall accuracy). A hybrid solution was found using manually selected aggressive keywords and automatically selected comments.

In the project *Virtual Aggression Barometer,* social scientists research overall daily aggression in the virtual space. The data are automatically gathered using 800 subjectively selected aggressive keywords[4] and coefficients from 0.2 to 1 subjectively assigned to each keyword. The public mood is determined every day by summing up the coefficients. A data analysis shows that this method tends to leave out important aggressive keywords and tends to include statically insignificant keywords. However, the keywords were useful to select aggressive and non-aggressive training data for this research.

The training data were obtained by choosing comments with the highest weight of the aggressive keywords (as aggressive samples) and the ones with no aggressive keywords as non-aggressive samples. The aggressive comment samples contained on average 10 aggressive keywords. The non-aggressive comment samples were revised not to include ambiguous texts. The training data were selected to contain 10 thousand samples on each class.

Three thousand manually classified comments (previously described) were used for the test data. These data were considered as a golden standard. Two transformation approaches on the training and test data were used to evaluate the connection between text normalization and sentiment analysis. First, the data were normalized using the tool discussed further and then lemmatized using a programming library *LVTagger* [2] and *Morphology* [3]. In the lemmatization process, insignificant word classes were filtered out (prepositions, conjunctions, particles, punctuation marks, residuals).

## 2. Normalization

The text normalization was performed using a rule based solution which analyzes variants created and then chooses the best one using n-gramms. The tool was originally created to correct OCR mistakes in historical Latvian texts [4]. Rules for variant generation and exact replacement were created (~5 thousand new rules) and the solution

[4] http://barometrs.korpuss.lv/?from=2013-07-14&to=2014-07-14&section=keywords

was adapted to correct mistakes using the statistically best probable variant by Latvian letter and word n-gramms (made on balanced contemporary Latvian corpus [5]). More features of the adapted version include language (English and Russian) recognition. The solution performs with 92% accuracy. The accuracy was determined using manually corrected texts – a golden standard.

## 3. Discussion and Results

Overall results (Table 1) of the sentiment analysis solution show that the aggressive and the non-aggressive comment distinction can be made with 6.2% drawback from the golden standard 78%. The results support the claim that the normalization improves the overall accuracy of the automatized sentiment analysis (by 2.4%). The results decrease if lemmatization is introduced. This can be explained by the sentiment information hidden in the inflection of the verbs. For example, if a verb is used in the second person then it might indicate aggression more often compared to when used in the first person.

**Table 1.** Overall results

| Variant | Feature count | Overall accuracy (%) |
|---|---|---|
| Original | 69,809 | 70.6 |
| **Normalized** | **63,251** | **72.2** |
| Filtered and lemmatized | 36,948 | 71.8 |

Both aggressive and non-aggressive comment recognition f-score (Table 2 and Table 3) has improved by the normalization; however, the aggressive comment recognition is still low. It could be improved by using more precise aggressive keyword list when automatically selecting aggressive and non-aggressive training sets. The system could be improved in a spiral development: by using keywords and coefficients calculated by the Naïve Bayes classifier to improve existing keyword list and then automatically gathering a larger aggressive and non-aggressive comment training sets.

**Table 2.** Overall results of aggressive comment recognition

| Variant | Feature count | Aggressive precision (%) | Aggressive recall (%) | Aggressive f-score (%) |
|---|---|---|---|---|
| Original | 69,809 | 26.8 | 38.0 | 31.5 |
| **Normalized** | **63,251** | **28.9** | **38.8** | **33.1** |
| Filtered and lemmatized | 36,948 | 28.5 | 38.8 | 32.9 |

**Table 3.** Overall results of non-aggressive comment recognition

| Variant | Feature count | Non-aggressive precision (%) | Non-aggressive recall (%) | Non-aggressive F-score (%) |
|---|---|---|---|---|
| Original | 69,809 | 85.3 | 77.6 | 81.3 |

| | | | | |
|---|---|---|---|---|
| **Normalized** | **63,251** | **85.7** | **79.4** | **82.4** |
| Filtered and lemmatized | 36,948 | 85.7 | 78.9 | 82.2 |

There is no possible exact comparison between the manually selected aggressive keyword search method and the aggressive and non-aggressive comment classification, but a visual comparison can be made (Figure 1). Both graphs match in the extremes, thus, confirming that the new method identifies the necessary trends. The new method extends the previous by analyzing all the words and statistically assigning the coefficients.

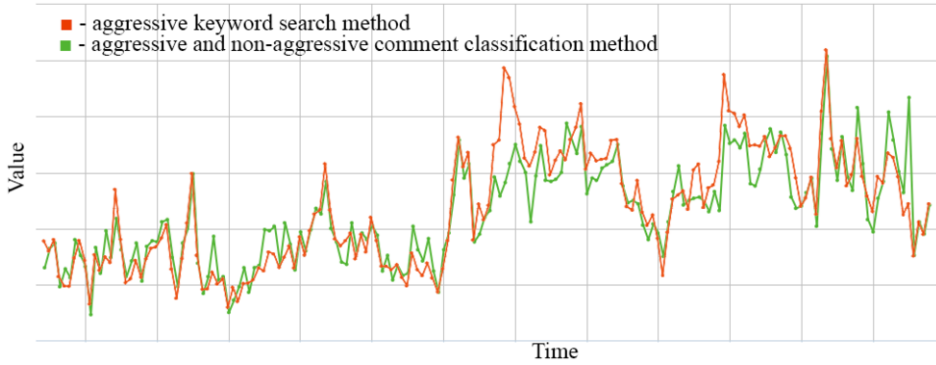


**Figure 1.** Visual comparison of aggression over time

## 4. Conclusion

Sentiment analysis is affected by the transliteration mistakes in contemporary online Latvian language. Normalizing texts improves overall results of the Naïve Bayes classifier for an aggressive and non-aggressive comment distinction by 2.4%. However, the sentiment analysis results could be improved by using the new data of the aggressive keywords to gather better training data.

## References

[1] Vryniotis, V. Developing a Naive Bayes Text Classifier in JAVA. Available: http://blog.datumbox.com/developing-a-naive-bayes-text-classifier-in-java/ [27.05.2014].

[2] Paikens, P., Pretkalniņa, L., Rituma, L. Morphological analysis with limited resources: Latvian example. In: Oepen, S., Hagen, K., Johannesse, J. (eds.). *Proceedings of the 19th Nordic Conference of Computational Linguistics*. 2013, pp. 267–277.

[3] Paikens, P. Lexicon-Based Morphological Analysis of Latvian Language. In: *Proceedings of 3rd Baltic Conference on Human Language Technologies (HLT 2007)*, Kaunas, 2007, pp. 235-240

[4] Pretkalniņa, L., Paikens, P., Grūzītis, N., Rituma, L., Spektors, A. Making Historical Latvian Texts More Intelligible to Contemporary Readers. In: *Proceedings of the Workshop on Adaptation of Language Resources and Tools for Processing Cultural Heritage Objects (LREC 2012)*, Istanbul, 2012, pp. 29–35.

[5] *Līdzsvarots mūsdienu latviešu valodas korpuss.* Available: http://www.korpuss.lv/ [22.05.2014.].