Styles with Benefits. The StyloMetrix Vectors for Stylistic and Semantic Text Classification of Small-Scale Datasets and Different Sample Length

Inez Okulska[0000-0002-1452-9840] and Anna Zawadzka[1]
[1] NASK National Research Institute, Warsaw, Poland
inez.okulska@nask.pl

Abstract. This paper offers a new approach to the problem of text classification of small-sized datasets and datasets with text samples of different lengths. StyloMetrix is a tool that allows for representing a text sample of any length with a linguistic vector of a fixed size – with 138 features. Each feature in the vector translates into a quantified, reproducible measure of an objective linguistic phenomenon – i.e., normalized statistics of chosen occurrence of morpho-syntactic or lexical relations. Thus, StyloMetrix vectors also show great potential in explainable AI. The StyloMetrix vector can serve as input data for classification algorithms, achieving high accuracy for relatively small-scale datasets with samples of varying lengths, compared to the semantic embeddings. Even though StyloMetrix does not encode any semantic information, the vectors (despite its name) also proved to be a valuable tool for content classification.

Keywords: Text Classification, Stylometry, Document Embeddings.

1 Vector Representations for Text Classification

1.1 Introduction

One of the challenges of text classification is the right choice of text representation, trading off the performance, resources, adjustments to the dataset or explainability. The most popular Word2Vec [8], Glove [10], or FastText [1] embeddings still amaze us with their universality for various tasks. However, the semantic relations between words depend strongly on the training particular corpus, and homonyms are represented by only one vector (no disambiguation). The BERT [2] embeddings are contextual and much more precise. However, even the pre-trained models often require extensive training data to be fine-tuned for a new task. Moreover, word embeddings present the problem of sample size, which needs to be equal or padded or trimmed, which is challenging for datasets with a wide range of sample lengths.

This paper proposes an extended, proprietary, stylometric vector representation obtained with the StyloMetrix tool: the StyloMetrix vectors. Besides resolving the issue of varying sample length (the vectors encode entire documents, no matter the size), they offer a different perspective on the represented text, i.e., a linguistic one. Not the "meanings" of the words are encoded, but the stylistic structure of the entire sample. StyloMetrix vectors have proved to have good performance not only for style, but

also for content classification – the vectors are sensitive to semantics, even for small-scale datasets with samples of varying lengths.

1.2 Related Work

Stylometric research typically tackles authorship attribution and semantic classification tasks. The focus is often on words frequencies or stopwords incidences [13][14]. For the Polish language, such a lexical approach is taken in the stylo package for R created by Eder and Rybicki [3], which offers text representation as vectors of relative frequencies of extracted n-grams. For reliable performance, texts need to be of 5000 tokens at minimum. Representing entire documents with fixed-size vectors has been proposed in Doc2Vec [7], i.e., an enhanced CBOW-like method based on Word2vec (semantic) word embeddings, and by Ryciak et al. l [12], who introduced word histograms that produce normalized vectors of 9000 features. It has shown great potential for unsupervised classification tasks of an extensive, unbalanced data set of patent documents. As for the morphosyntactic approach, the Websty tool [11] for the Polish language offers document comparison by manual feature selection from computed distributional metrics. However, it does not cover syntactic structures.

2 The Method (The StyloMetrix Vectors)

StyloMetrix is a tool that allows for representing a text sample of any length with a linguistic vector of a fixed size – 138 features. Each feature in the vector translates into a quantified, reproducible measure of the chosen occurrence of morpho-syntactic or lexical relations. The metrics so far include: a) POS classification, b) grammatical forms of nouns, pronouns, c) comparative degrees in adjectives and adverbs, d) verbs in grammatical persons, aspect, tense, mood, in participle form, e) type-token ratio, f) incidence of content and function words, g) incidence of names and personal names, h) volume of 1%/5% of most common types for text, i) psycholinguistic metrics related to an experimental study by Imbir [4], j) incidence of declarative, interrogative and exclamative sentences, k) participation of words in nominal phrases and modifiers, l) sentence starts, syllable length and m) the presence of categorical affixes. The metrics are normalized, i.e. their values are relative to the total number of words (alphanumerical tokens) in a sample, always providing a value in the range [0, 1]. They relay on morphosyntactic information in the spaCy model for the Polish language [15].

The StyloMetrix vector can serve as input data for classification algorithms, achieving high accuracy for relatively small-scale datasets with text samples of unbalanced length. The encoding does not transmit any semantic information as it is usually not subject to interest in stylometric experiments. However, it also proved to be a valuable tool for content classification. The example of 10 metrics computed for four excerpts from the dataset used further in the first experiment are shown in Table 1. To validate the metric values in the Table 1, we present two from the analyzed excerpts, one from non-professional erotic stories (A) and non-professional neutral short stories (B), accordingly. The trends presented on these random samples mostly represent their classes, hence we decided to use them for underlining the possible contrast.

It seems that a certain combination of linguistic choices (e.g. the range of vocabulary, verbs over nouns, the use of contemporary transgressives, or 1st person singular) can denote semantic aspects specific to a text genre, i.e. erotic stories in this case.

Table 1. Chosen metrics for random excerpts from each category

Class	Sample 1 n-prof erotic	Sample 2 n-prof neutral	Sample 3 non-fiction	Sample 4 pop lit
Verbs	**0,27**	0,19	0,18	0,23
Nouns	0,27	0,32	0,42	0,37
Pronouns	**0,20**	0,12	0,03	0,07
1st person singular verbs	**0,17**	0,09	0	0
3d person singular verbs	0,03	0,05	0,13	0,20
Verbs in past tense	**0,17**	0,12	0,01	**0,20**
Contemporary transgressives	**0,07**	0	0	0
3rd person singular pers. pronouns	**0,10**	0	0,01	0,03
Tokens covering 1% of most common types	**0,17**	0,09	0,11	0,10
Words in a nominal phrase	0,30	0,63	0,73	0,67

A: *Uśmiechnęłam się do niego. Ponownie przeciągnęłam językiem po czubku penisa, obserwując, jak Adam przymyka powieki. Rozchyliłam wargi i naprowadzając dłonią penis, powiodłam nim po nich. Wreszcie wsunęłam go do ust.*

B: *W międzyczasie zjadłem obiad, później założyłem elegancką koszulę. Z biegiem czasu jednak robiło się niepokojąco cicho. Postanowiłem dowiedzieć się od współlokatorki Alicji, jak z przygotowaniem na imprezę i o której godzinie ma zjawić się Magda. Jednak nie uzyskałem odpowiedzi od razu!*

3 The Experiments

Two classification tasks were tackled: text genres with or without adult content and sentiment analysis, both on small-scale datasets. We used StyloMetrix vectors as input to Random Forests. Text genre classification has also been compared with the results of a RNN with BiLSTM layer and the BERT model.

We gathered four classes of samples of varying lengths (short stories, novels, reportages), including non-professional erotic stories and non-professional neutral stories, non-fiction, and pop literature (for statistics see Table 2). The data was split 80:20 for training and test. Each text sample was represented with the StyloMetrix vector of size 138. For the LSTM and BERT models, we split the data into text chunks of 250 tokens, with punctuation. Then we represented the chunks accordingly with vectors of size 100 (Polish Word2vec)[5] and 1024 (HerBERT large case)[9].

Table 2. Dataset statistics for the supervised text classification experiment

Class	n-prof erotic	n-prof neutral	non-fiction	pop lit	sum
Sample count	263	300	**556**	189	1308
Min sample length	215	304	**42**	543	-
Max sample length	5551	9825	40363	**130220**	-
Chunks count	801	1998	1108	**5234**	**9141**

The first experiment, i.e. supervised text classification with Random Forests and StyloMetrix vectors representing the text genres dataset (1308 samples) yielded the mean accuracy of 93.5%, offering balanced scores across classes as shown in Table 3. Classification using deep learning models on the same data resulted in poor performance, however the adjusted dataset of chunks (creating a bigger data set of 9141 samples) yielded 83.5% accuracy for LSTM and 84.25% for the HerBERT model.

Table 3. Supervised four text genres classification results for different word/text representation

Accuracy per class	n-prof erotic	n-prof neutral	non-fiction	pop-lit
RF with StyloMetrix (original samples)	**0.91**	**0.92**	**1.0**	0.91
BiLSTM with Word2vec (70 ep., 250-token chunks)	0.85	0.79	0.81	0.93
HerBERT large (200 epochs, 250-token chunks)	**0.91**	0.67	0.74	**0.95**

For the second experiment we used a subset of 607 samples of opinions in 2 domains (school and medicine) from the PolEmo 2.0 Sentiment Analysis Dataset [6]. We preprocessed the data to correct frequent typing errors in verbs (eg. *popatrzył em*) and used StyloMetrix to produce a separate NumPy matrix file (.npy) for each group[1].

Table 4 presents the results of the second experiment: classification of samples belonging to 4, 3, and 2 sentiment groups on the small subset and the full set. The ambivalent class turned out to be indistinguishable from positive samples in style in the current experimental setup. That is why we decreased the number of the classes – first excluding only the ambivalent category, and then the neutral as well.

The results indicate the best classification level for binary classification using the StyloMetrix data with a very simple classifier. It shows that sentiment can be traced in stylistic features, not only the semantic ones (like in most of the related approaches).

Table 4. Supervised sentiment classification with StyloMetrix vectors

Accuracy per class	Positive	Ambivalent	Neutral	Negative	Mean acc
4 classes – subset	0.81	0	0.54	0.58	0.48
3 classes – subset	0.88	-	0.62	0.76	0.75
2 classes – subset	**0.89**	-	-	**0.86**	0.86

4 Conclusions

The presented StyloMetrix vectors performed well on small-scale datasets (about 200-300 samples per class) with texts of different lengths, obtaining good results in classifying text genres (including separating the adult content from neutral ones) and sentiment in clients reviews. The results confirm our initial claims: StyloMetrix vectors provide novel, rich and qualitative representation of textual data that can even be

[1] The StyloMetrix vectors for this particular dataset may be downloaded from https://github.com/ZILiAT-NASK/Datasets/tree/main/StyloMetrix_PolEmo.

sensitive to semantics. We are thrilled to find out about further usages of this data as we extend the scope of metrics and try different experimental scenarios.

References

1. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the association for computational linguistics 5, 135-146 (2017).
2. Devlin, J., Chang, M. W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018).
3. Eder, M., Rybicki, J., Kestemont, M.: Stylometry with R: a package for computational text analysis. The R Journal 8(1), 107-121 (2016).
4. Imbir, K. K.: Affective norms for 4900 Polish words reload (ANPW_R): Assessments for valence, arousal, dominance, origin, significance, concreteness, imageability and, age of acquisition. Frontiers in psychology 7, Article 1081 (2016).
5. Kędzia, P., Czachor, G., Piasecki, M., Kocoń, J.: Vector representations of polish words (Word2Vec method). CLARIN-PL digital repository, http://hdl.handle.net/11321/327, last accessed 14/04/2022. (2016).
6. 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), pp. 980-991. Association for Computational Linguistics, Hong Kong (2019).
7. Lau, J. H., Baldwin, T.: An empirical evaluation of doc2vec with practical insights into document embedding generation. arXiv preprint arXiv:1607.05368 (2016).
8. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013).
9. Mroczkowski, R., Rybak, P., Wróblewska, A., Gawlik, I.: HerBERT: Efficiently pretrained transformer-based language model for Polish. arXiv preprint arXiv:2105.01735 (2021).
10. Pennington, J., Socher, R., Manning, C. D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543. Associatio for Computational Linguistics, Doha (2014).
11. Piasecki, M., Walkowiak, T., Eder, M.: Open stylometric system WebSty: integrated language processing, analysis and visualization. Computational Methods in Science and Technology 24(1), 43-58 (2018).
12. Ryciak, N., Chrabąszcz, M., Bartoszuk, M., Classification of patent applications, an oral seminar held at the Institute of Computer Science, Polish Academy of Sciences (IPI PAN) available online as recorded video (www.youtube.com/watch?v=L8RRx9KVhJs) (2020)
13. Rybicki, J., Heydel, M.: The stylistics and stylometry of collaborative translation: Woolf's Night and Day in Polish. Literary and Linguistic Computing 28(4), 708-717 (2013).
14. Savoy, J.: Machine Learning Methods for Stylometry. Springer, Cham (2020).
15. Tuora, R., Kobylinski, Ł.: Integrating Polish language tools and resources in Spacy. In: Proceedings of PP-RAI 2019 Conference, pp. 210-214. Department of Systems and Computer Networks, Faculty of Electronics, Wroclaw University of Science and Technology, Wrocław (2019).
16. Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., Xu, B.: Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers), pp. 207-212. Association for Computational Linguistics, Berlin (2016)