Detrended fluctuation analysis of sentiment patterns in literary texts

Łukasz Gaża
Faculty of Computer Science and Telecommunications, Cracow University of Technology
31-155 Kraków, Poland
E-mail: lukasz.gaza@pk.edu.pl

Paweł Oświęcimka
Institute of Nuclear Physics, Polish Academy of Sciences
31-342 Kraków, Poland
Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University
30-348 Kraków, Poland

Stanisław Drożdż
Institute of Nuclear Physics, Polish Academy of Sciences
31-342 Kraków, Poland Faculty of Computer Science and Telecommunications, Cracow University of Technology
31-155 Kraków, Poland

Marek Stanuszek
Faculty of Computer Science and Telecommunications, Cracow University of Technology
31-155 Kraków, Poland

KEYWORDS

Sentiment analysis; Literary texts; Detrended Fluctuation Analysis, Hurst exponent

ABSTRACT

Temporal correlations of the sentiment content in consecutive sentences are studied based on a large corpus of the world-famous literary texts in four major European languages (English, French, German, and Spanish). For quantifying the related characteristics in terms of the Hurst exponents the Detrended Fluctuation Analysis (DFA) is employed. The results obtained provide a clear indication for the existence of persistent correlations as revealed by the Hurst exponents significantly larger than 1/2 in all the four languages studied. An interesting result that requires a deeper study is the identified fact that in many texts the DFA indicates two different regimes of scaling and thus two different Hurst exponents. In those cases - quite universally - the cross-over occurs at the scale corresponding to about 200 sentences and the Hurst exponents at the scales above this value are even larger which indicates the presence of stronger long-range temporal correlations than those at the shorter scales.

I. INTRODUCTION

The aim of the presented study is to examine the application of DFA to sentiment analysis of literary texts. The topic of sentiment analysis is important due to the growing demand for sentiment classification tools. Also, sentiment analysis of literary texts alone is important to answer the question if there are any characteristics specific to some particular authors or specific to texts created by humans versus text generated by large language models.

Our main contributions are:

- analysis of long-range sentiment correlations in literary texts;
- application of DFA to sentiment analysis.

The paper is organized as follows. Section (II) is describing the sentiment analysis technique and its applications. Section (III) is describing the process of assessing the distribution of sentiment within the texts and the methods of calculating the correlation between sentences. It also describes how the simulations were performed. Section (IV) discusses the outcomes of the simulations. Indications were obtained that there is a positive long-range correlation between sentences' sentiment in literary texts. The paper ends with section (V), which contains conclusions based on the conducted simulations and obtained results. Ideas for the future work are also discussed there.

II. RELATED WORK

Sentiment analysis is a text classification technique that allows the examination of emotional charge and its distribution within a text [6][13]. Among other areas, sentiment analysis is usually employed in: a) Search and classification of emotionally-charged words, b) Harvest of insights into the opinions of customers on products or services, especially useful for business and social sciences, and c) Analysis of emotions in literary texts. The related research topic comprises tools used to generate sentiment (tools allowing scoring texts or individual words to be able to calculate sentiment based on those). In the case of this kind of tools, one can analyze the influence of responders' age, gender, or education level on the way they score the words [12][15]. The research can also examine the influence of language on scoring.

The characteristics of the sentiment may differ depending on the context in which the text is analyzed.

Communications of the ECMS, Volume 37, Issue 1,
Proceedings, ©ECMS Enrico Vicario, Romeo Bandinelli,
Virginia Fani, Michele Mastroianni (Editors) 2023
ISBN: 978-3-937436-80-7/978-3-937436-79-1 (CD) ISSN 2522-2414

When the goal is the analysis of the customers' experience regarding some product or service, the source texts are usually short, like social media posts or press releases [9]. The sentiment in this kind of source is usually constant across the entire text because it depends on the author's experience regarding a given product or service [10].

Another research area is the analysis of literary texts, in which the focus is set on understanding how the sentiment evolves over time. In [14] it was shown that all literary texts could be characterized by six basic shapes. The application of sentiment analysis techniques to literary texts may help to provide insights in the perception of the book. Another goal might be to learn about the mental state of the author.

The application of Time-Series techniques to the analysis of literary texts has gained traction over time in the research community. In [8] it was, for instance, shown that there are long-range correlations in the sentence length variability. Still, whether some analogous kind of correlation exists for sentiment series in texts is an open question. The present work takes an approach to address that question. The literary texts are thus treated as time-series sequences in which each element of the sequence reflects the sentiment contents of consecutive sentences. Using the commonly recognized Detrended Fluctuation Analysis (DFA) we then ran multiple simulations in order to reveal the character of correlations of those sentiments for the corpora of literary texts in four European languages.

III. METHODS

Dictionaries used in the research

The sentiment score was computed for each sentence based on the sentiment scores of each individual word and stored in dictionaries. The dictionaries of sentiments of individual words were downloaded from the hedonometer page [2]. The original dictionaries were created using Mechanical Turk [1]. A set of words from different corporas (specific for a given language) was manually given a score by 50 native speakers - they rated how they felt in response to individual words. The score was between 1 (most negative emotion) and 9 (most positive emotion). In total 5 million individual human assessments were performed. The result of that process was a list of 10000 most popular words and their sentiments calculated as the average of individual scores [7]. As part of the research described in this article, the sentiment was scaled down to the range from -4 (most negative emotion) up to 4 (most positive emotion). It was done to decrease the influence of neutral words on the sentiment of the sentence. Because the dictionaries consisted of 10000 of the most popular words for each language, not all the words from the literary texts existed in those dictionaries (approximately 30% of the words from the texts did not exist in the dictionaries). Those words were given a score of 0, so they were considered neutral words.

The distribution of sentiment score of individual words in the dictionaries is shown in Figure 1.

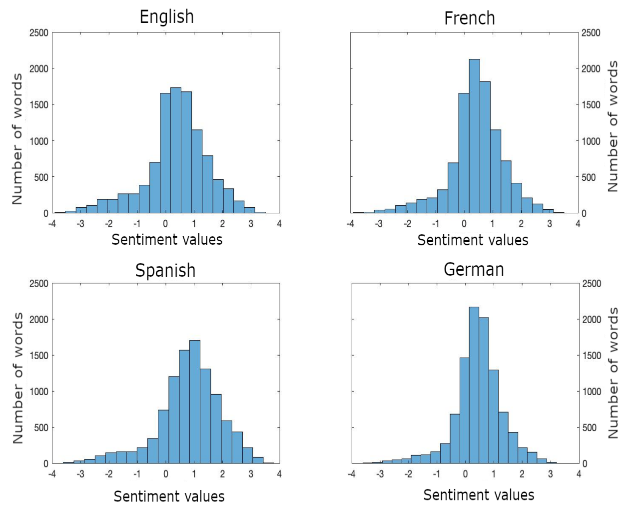

Fig. 1: The distribution of sentiment of individual words in the dictionaries.

The resulting dictionaries presented a set of common attributes for all the languages under consideration. First of all, the average sentiment of each of the dictionaries was positive. Moreover, distributions of the dictionaries were left-skewed. It is in line with the thesis from article [7], where it was indicated that human languages exhibit a clear positive bias which is a big data confirmation of the Pollyanna hypothesis[5].

Literary texts used in the research

The literary texts written in English, French, German, and Spanish have been selected for the present study. The following criteria were used for the selection of the literary texts in this work: a) they belong to the world's major languages, b) the dictionaries mentioned in the previous section are available for those four languages and c) a large number of literary texts in those languages is freely available. For each language, we prepared a set of 100 texts. All texts are composed of, at least, 3000 sentences. A sentence is defined as a sequence of words starting with a capital letter and ending in a full stop. For the correlated series, as the ones to be studied here, such a lower bound on the number of sentences is required to obtain statistically reliable results. The distribution of sentiment in a set of books in each of the mentioned languages is shown in Figure 2.

Assigning sentiment content to a sentence

During the preprocessing phase, each literary text was split into sentences and the stop words were removed, as they comprise the most common non-emotional words. In the next step, we computed the sentiment score s_i for the ith sentence as the sum of sentiment scores for each individual word s_w divided by a number of all words in the sentence l_{sen}:

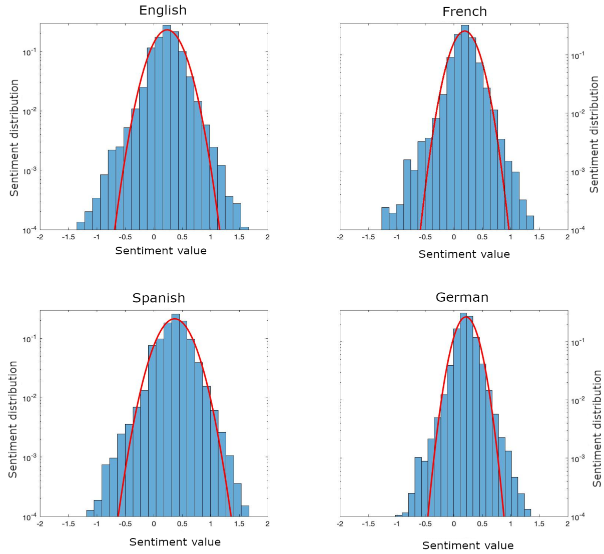

Fig. 2: Distribution of sentiment in a set of books in four languages. Redline is the Gaussian distribution.

$$s_i = \frac{\sum_{w=1}^{l_{sen}} s_w}{l_{sen}} \tag{1}$$

Not all words occurring in the sentences existed in the source dictionaries because the dictionaries contain only 10000 of the most popular words for each language. Those words not existing in the dictionaries were given $s_w = 0$, so they were treated as neutral words.

There were some other possibilities for calculating the sentiment of the sentence s_i considered.

The first one was using in the denominator the number of scored words in the sentence l_{score} (number of words from the sentence which existed in the source dictionary) rather than the number of all words. This method was not giving significantly different results than the previous method because it was just decreasing the denominator of the formula $s_i = \frac{\sum_{w=1}^{l_{sen}} s_w}{l_{score}}$.

Another considered method was to calculate the sentence sentiment s_i as sum of sentiments of individual words: $s_i = \sum_{w=1}^{l_{sen}} s_w$ That way of assigning sentiment to a sentence results in a much wider range of sentiment values, especially for long sentences as it involves the arbitrariness related to the length of a sentence.

Having sentence sentiment s_i we constructed series $S = \{s_i\}|i = 1, .., N$ where N is number of the sentences in the text.

Detrended Fluctuations Analysis

DFA is described in [11] as a method of determining the scaling behavior in time series x_i of i=1,...,N of equidistant measurements. The aim is to calculate correlations between values x_i and x_{i+s} for different time scales s. The DFA method consists of four steps. In the first step the profile $Y(i) = \sum_{k=1}^{i} x_k - < x >$ of records x_i is determined. In the second step, the profile $Y(i)$ is divided into $N_s = \lfloor N/s \rfloor$ non-overlapping segments of length s. To avoid disregarding any segments the same procedure is performed starting from the end of the record resulting in $2N_s$ segments. In the third step, the local trend of each segment is determined using the least-squares method. The detrended data of the segment $Y_s(i)$ is defined as the difference between the original time series and the estimated trends $Y_s(i) = Y(i) - p_\nu(i)$, where $p_\nu(i)$ is the fitting polynomial in the ν-th segment. In the next step the variance of each of the $2N_s$ segments of the detrended time series $Y_s(i)$ is calculated:

$$F_s^2(\nu) = \frac{1}{s} \sum_{i=1}^{s} Y_s^2[(\nu - 1)s + i]. \tag{2}$$

Finally, the average over all the $2N_s$ segments and its root square is taken to obtain the DFA fluctuation function $F(s)$:

$$F(s) = [\frac{1}{2N_s} \sum_{\nu=1}^{2N_s} F_s^2(\nu)]^{\frac{1}{2}} \tag{3}$$

In this work, the second-order polynomial is used for detrending. The presence of long-range correlation manifests itself in the power-law dependence of fluctuation functions $F(s)$ as follows:

$$F(s) \propto s^H \tag{4}$$

where H is the Hurst exponent. $H = 0.5$ indicates the lack of correlations, $H < 0.5$ anty persistent correlations, and $H > 0.5$ the persistent correlations.

Simulation tool

All the text processing was performed using Python code and the nltk library [3]. The simulations were performed on a personal computer using PyCharm Integrated Development Environment [4]. The below code presents the algorithm for assigning a sentiment score to a sentence:

```
def score_slice(scored_input_words_df, filter_fun):

    scores_by_words =
         scored_input_words_df.get_scores_by_words()

    def
         score_slice_as_avg_of_scored_words_helper(slice):
        total_score = 0.0
        count = 0

        for word in slice:
            word = word.lower()
            if word in scores_by_words:
                score = scores_by_words[word]
                if filter_fun(score):
                    total_score += score
                    count += 1

        if count == 0:
            return 0
        return total_score / count

    return score_slice_as_avg_of_scored_words_helper

def score(self, slice_size, scored_input_words_df):
    sentences = self.book.get_sentences()
```

```
    self.scores = list(map(analysis.scoring.score_slice(
    scored_input_words_df,
    analysis.scoring.TRUE_FILTER), sentences))

    return self.scores
```

The Hurst exponent was calculated using code written in Matlab.

IV. RESULTS

Statistical properties of sentiment series

Sentiment series for representative examples of literary texts in English, Spanish, French, and German are shown in Figure 3. The figures also show the per-sentence distribution of sentiment scores for those texts. As can be seen, the average sentiment of those texts is greater than 0 meaning it's positive. It is in line with the analysis of a larger set of texts in mentioned languages [7][5].A set of relevant statistical parameters of the distribution of sentiment scores for the collection of analyzed books is shown in Table I. In any of the analyzed books the sentiment score of the sentences doesn't take extreme values - the minimum is greater than -2 and the maximum is lower than 2.2.
There is no major difference between analyzed languages. The distribution is skewed left for English, French, and Spanish. For German the distribution is symmetric. The kurtosis value is higher than for the Gaussian distribution, the kurtosis for the Gaussian distribution is 3, which indicates heavier tails of the distribution than in the Gaussian case.

Language	mean	stddev	min	max	skewness	kurtosis
English	0.23	0.23	-1.89	2.2	-0.226	6.50
French	0.19	0.20	-1.81	1.89	-0.269	8.06
German	0.21	0.17	-1.66	1.67	0.017	7.54
Spanish	0.36	0.25	-1.91	2.07	-0.22	5.01

TABLE I: Statistical parameters of the distribution of sentiment in a set of books in a given language.

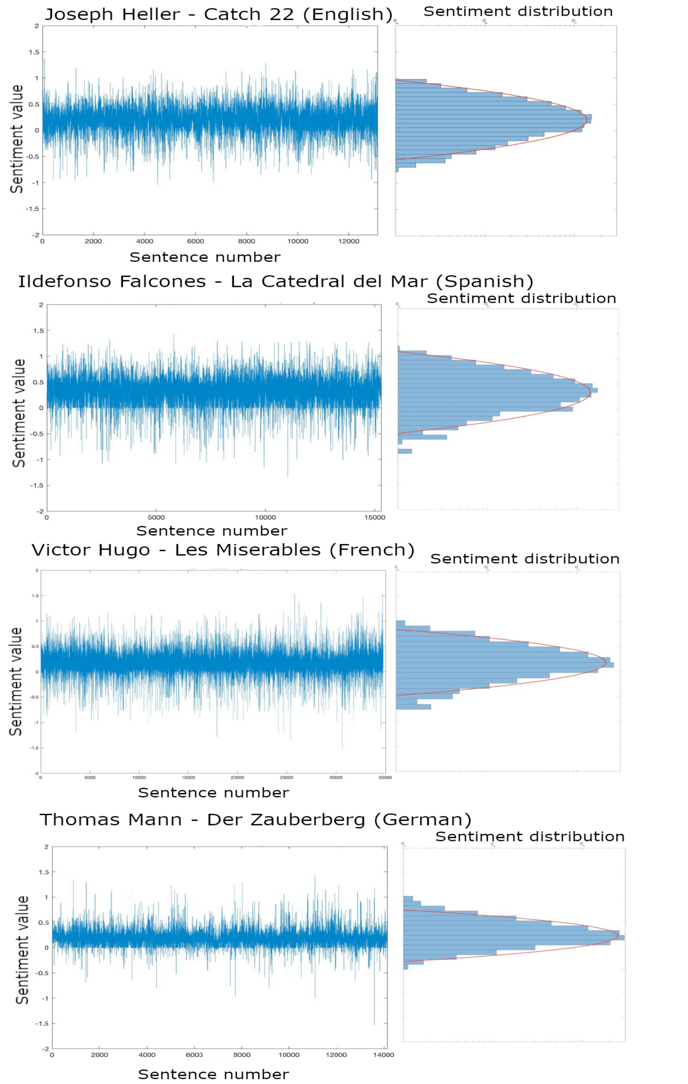

Fig. 3: Sentiment series for representative examples of books in four languages. The figures on the right side illustrate the per-sentence distributions of the sentiment score for the series under consideration. The deviation with respect to the Gaussian distribution, drawn as the red line, towards longer-tailed distributions, is clearly visible.

Determining the Hurst exponents

Figure 4 shows the sample fluctuation functions, as defined in 3. As it can be seen, the correlations between sentences with a distance lower than 200 sentences scale differently than longer-range correlations - there is a crossover point after approximately 200 sentences. This means that, after 200 sentences, the sentiment of the literary text changes.

The type of those correlations can be described by Hurst's exponent. There were simulations done to calculate the Hurst exponent for a set of literary texts (100 texts for each language). The values of the Hurst exponent for short-range correlation (a distance shorter than 200 sentences) and long-range correlation (a distance greater than 200 sentences) are shown in Table II. The Hurst exponent for the long-range correlations is larger than for the short-range correlations - there are some short-range trends that are persistent and on top of them there are few long-range trends with even higher persistence. The distribution of the Hurst exponent for the analyzed texts is presented in Figure 5.

	Short-range (<200 sent.)		Long-range (>200 sent.)	
Language	mean(H)	stddev(H)	mean(H)	stddev(H)
English	0.61	0.04	0.70	0.06
French	0.60	0.03	0.66	0.05
German	0.64	0.04	0.65	0.04
Spanish	0.60	0.03	0.69	0.06

TABLE II: Hurst exponent for all analyzed texts.

As it can be seen, the mean value of the Hurst exponent is approximately 0.6, which means the sentiments are persistent - there is a positive correlation for

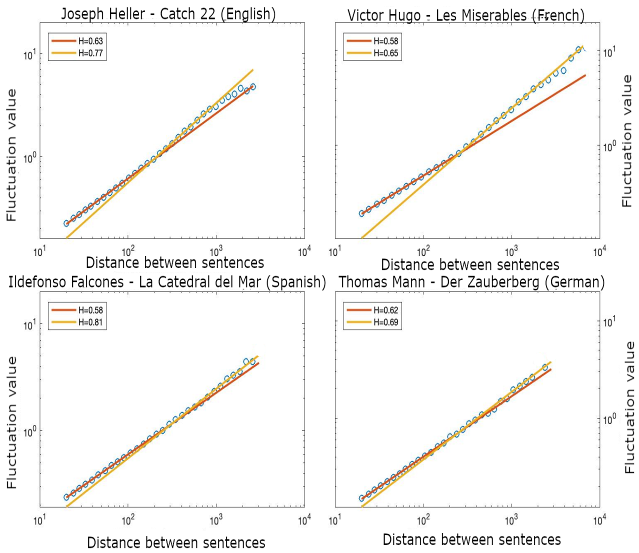

Fig. 4: DFA fluctuation functions for exemplary texts.

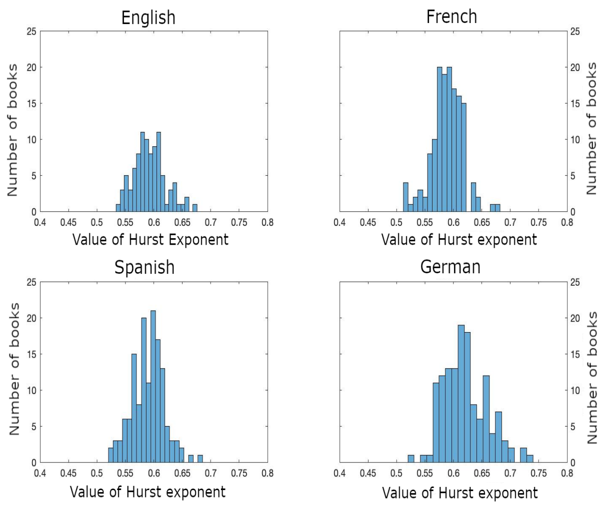

Fig. 5: Distribution of the Hurst exponent for all analyzed texts.

the sentiment. Moreover, Hurst exponent for all of the books is higher than 0.5 - the range is usually between 0.52 and 0.68. German books present even higher values for the Hurst exponent.

Surrogate tests

In order to verify the statistical significance of the present DFA analysis, surrogate tests are required. In the present case of sentiment series, the surrogate series were created by drawing randomly (with equal probability) words from the dictionaries by preserving the lengths of original texts, the lengths of individual sentences, and the number of words existing in the dictionary for the given language for each sentence. Examples of sentiment series for those surrogates are shown in Figure 6 and their distribution of sentiment in Figure 7.

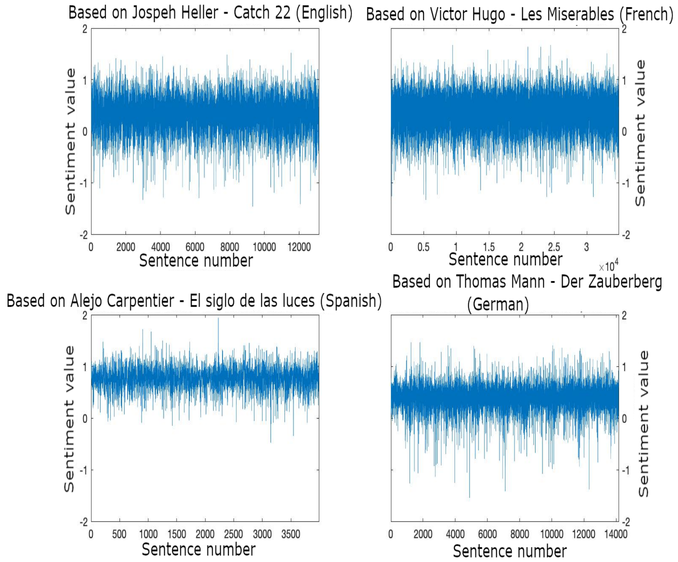

Fig. 6: Examples of sentiment series for surrogates created by randomly drawing words from the dictionaries with the statistical characteristic of the original books preserved.

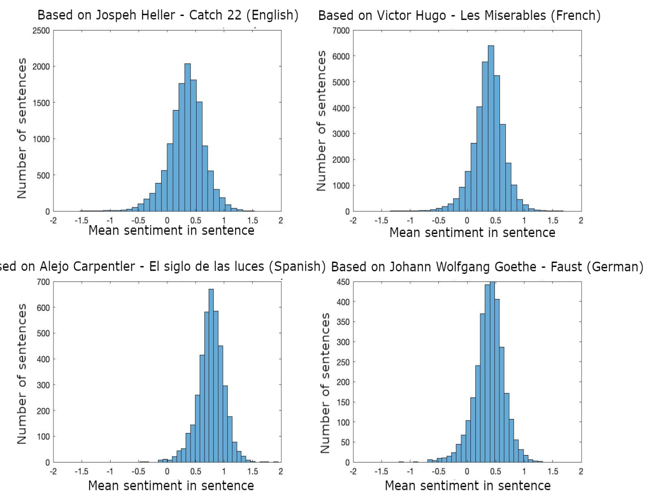

Fig. 7: Sentiment distribution for surrogates created by randomly drawing words from the dictionaries with the statistical characteristic of the original books preserved.

Calculating Hurst exponent (H) for those surrogates reveals that the resulting exponents are close to 1/2 - there is thus no persistence of sentiment for such surrogate texts. The representative fluctuation functions for this kind of surrogate are shown in Figure 8.

Another obvious type of surrogate is obtained by shuffling the order of sentences. A comparison of Hurst exponent between both types of surrogates is shown in Table III.

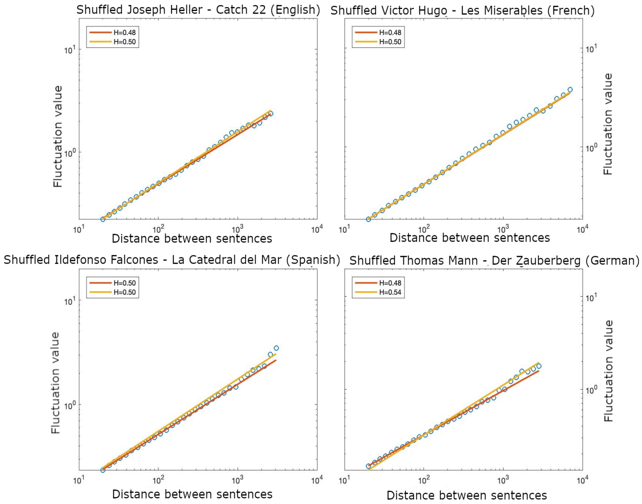

Fig. 8: Examples of fluctuation functions in different languages for books generated based on real books.

	Generated texts		Shuffled texts	
Language	mean(H)	stddev(H)	mean(H)	stddev(H)
English	0.54	0.044	0.52	0.04
French	0.54	0.03	0.51	0.03
German	0.55	0.03	0.52	0.03
Spanish	0.57	0.04	0.53	0.04

TABLE III: Hurst exponent for randomly generated and shuffled surrogates.

V. CONCLUSIONS

Based on the DFA methodology, the exploratory study of correlations in the sentiment content through sentences in literary texts representing four major Europen languages (English, French, German, and Spanish), as presented in this work, it indicated a quite universal presence of long-range temporal correlations of persistent character in the distribution of sentiment score of the sentences as revealed by the Hurst exponents significantly larger than 1/2. In many cases, such correlations involve even two components with the scale cross-over corresponding to distances of about 200 sentences in separation. For distances larger than these values the correlations preserve their persistent character and, somewhat unexpectedly, get stronger as reflected by even larger values of the corresponding Hurst exponents. This effect of cross-over may reflect the existence of global patterns (six types exhausting all possibilities) of sentiment as postulated in [14]. At the larger scales such long-term global patterns may constitute the leading factor carrying correlations. The statistical significance of the above results has been validated by performing tests on the corresponding surrogate series with correlations destroyed by appropriate randomization. For such surrogate series, the DFA analysis adopted here results in the Hurst exponents close to 1/2, thus reflecting the lack of correlations as it should.
To sum up the present analysis indicates an interesting direction of research. The future work will cover a more systematic study that would allow more firm conclusions on what is universal and what system is specific in the sentiment patterns in literature. In particular, we plan to apply a multifractal generalization of DFA (MFDFA) as well as we are going to study cross-correlations (possibly of the fractal character) between the time series of sentence length and the corresponding series of the normalized sentiment content. Finally, we are going to analyze the difference between the languages in the context of the long-range correlations and the cross-over.

References

[1] Amazon mechanical turk, https://www.mturk.com/.
[2] Hedenometer, https://www.hedonometer.org/.
[3] Natural language toolkit, https://www.nltk.org/.
[4] Pycharm integrated development environment, https://www.jetbrains.com/pycharm/.
[5] J. Boucher and C. Osgood. The pollyanna hypothesis. *Journal of Verbal Learning and Verbal Behavior*, 8(1-8), 1969.
[6] M. D. Devika, C. Sunitha, and A. Ganesh. Sentiment analysis:a comparative study on different approaches. *Procedia Computer Science*, 87:44–49, 2016.
[7] P. S. Dodds, E. M. Clark, S. Desu, M. R. Frank, A. J. Reagan, J. R. Williams, L. Mitchell, K. D. Harris, I. M. Kloumann, J. P. Bagrow, K. Megerdoomian, M. T. McMahon, B. F. Tivnan, and C. M. Danforth. Human language reveals a universal positivity bias. *PNAS*, 112(8), 2015.
[8] S. Drożdż, P. Oświęcimka, A. Kulig, J. Kwapień, K. Bazarnik, I. Grabska-Gradzińska, J. Rybicki, and M. Stanuszek. Quantifying origin and character of long-range correlations in narrative texts. *Information Sciences*, 331:32–44, 2016.
[9] Z. Drus and H. Khalid. Sentiment analysis in social media and its application: Systematic literature review. *Procedia Computer Science*, 161:707–714, 2019.
[10] R. S. Jagdale, V. S. Shirsat, and S. N. Deshmukh. Sentiment analysis on product reviews using machine learning techniques. *Advances in Intelligent Systems and Computing*, 768, 2018.
[11] J. W. Kantelhardt1, E. Koscielny-Bunde1, H. H. A. Rego, S. Havlin, and A. Bunde. Detecting long-range correlations with detrended fluctuation analysis. *Physica A: Statistical Mechanics and its Applications*, 295(3–4):441–454, 2001.
[12] S. Kumar, M. Gahalawat, P. P. Roy, D. P. Dogra, and B.-G. Kim. Exploring impact of age and gender on sentiment analysis using machine learning. *Electronics*, 9(374), 2020.
[13] R. Prabowo and M. Thelwall. Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2):143–157, 2009.
[14] A. J. Reagan, L. Mitchell, D. Kiley, C. M. Danforth, and P. S. Dodds. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Sci.*, 5(31), 2016.
[15] S. Suman and S. Urolagin. Age gender and sentiment analysis to select relevant advertisements for a user using cnn. *Data Intelligence and Cognitive Informatics*, 2022.

AUTHOR BIOGRAPHIES

ŁUKASZ GAŻA He received his M.Sc. in the field of Applied Physics with Computer Modelling at the Tadeusz Kosciuszko Cracow University of Technology. Since 2019 he is a Research and Teaching Assistant at the Tadeusz Kosciuszko Cracow University of Technology. His e-mail address is lukasz.gaza@pk.edu.pl.

STANISŁAW DROŻDŻ is a professor of physics and head of the Complex Systems Theory Department at the Institute of Nuclear Physics (INP), Polish Academy of Sciences, and professor of computer science at the Cracow University of Technology. He received an M.S. (1978) and Ph.D. (1982) in physics, both from Jagiellonian University in Kraków, and a DSc (1988) in theoretical physics from INP. In 1994, he received the Polish state title of professor. His long-term scientific stays abroad include Forschungszentrum Juelich and Bonn University, Germany and the University of Illinois at Urbana-Champaign, USA. His research interests and activities include the quantum many-body problem, nonlinear dynamics, general aspects of complexity, brain-research-related issues, dynamics of financial markets including cryptocurrencies, and quantitative linguistics. Currently he also serves as Editor or Editorial Board Member for Physica A, Entropy, and for Chaos, Solitons and Fractals.

PAWEŁ OŚWIĘCIMKA Assoc. Prof., received his M.S. degree in physics from the AGH University of Science and Technology, Kraków, in 2001, and Ph.D. and D.Sc. in physics from the Institute of Nuclear Physics (INP), Polish Academy of Sciences, Kraków, in 2005 and 2015, respectively. Since 2016, he has been Associate Professor in the Complex Systems Theory Department at INP. He is the leader of the research group at the Jagiellonian University in the project on bio-inspired artificial neural networks, and he is a lecturer at the Cracow University of Technology. His current research and multidisciplinary collaborations involve using the multifractal approach to Computational Neuroscience, Dynamics of Financial Markets, and Quantitative Linguistic.

MAREK STANUSZEK He received his M.SC in the field of Theoretical Mechanics at the Faculty of Civil Engineering., Cracow University of Technology. Then he received his Ph. D. in the field of Computer Methods and Application based on the thesis: Computer Modeling of Flexible Structures with Finite Element Method. Finally, in 2005 he completed his habilitation thesis based on the work: Modeling of Flexible Structures with Unilateral Boundaries. Since 2005 he is the University Professor in the field of Computer Methods and Information Science of the Cracow University of Technology.