ISSN 2007-9737

Identification of Suicidal Tendencies of Individuals Based on the Quantitative Analysis of their Internet Texts

Tatiana A. Litvinova[1,2,3], Pavel V. Seredin[1,3,4], Olga A. Litvinova[2,3], Olga V. Romanchenko[5]

[1] Voronezh State University, Voronezh,
Russia

[2] Voronezh State Pedagogical University, Voronezh,
Russia

[3] Scientific Research Centre “Kurchatov Institute”, Moscow,
Russia

[4] Benemerita Universidad Autonoma de Puebla, Puebla,
Mexico

[5] Plekhanov Russian University of Economics, Moscow,
Russia

{centr_rus_yaz, olga_litvinova_teacher, ghjd1}@mail.ru, paul@phys.vsu.ru

Abstract. Even though suicide is one of the top three causes of young people’s deaths, no reliable methods of identifying suicidal behavior have been developed. One of the promising directions of research is quantitative analysis of speech. It is nowadays common to process texts by suicidal individuals (mostly suicidal notes or literary texts by famous people, e.g., poets, writes, etc.) and texts by individuals from a control group using software (mostly LIWC) and to design models for classifying texts as those by suicidal individuals or not. This kind of analysis has been mainly performed for English texts that generally have a number of restrictions due to their linguistic nature. The authors are the first to attempt to design a mathematical model to classify texts as those by suicidal or nonsuicidal individuals using numerical values of linguistic parameters as features. Texts (blogs by young people who committed suicides, similar in both genre and topic, to those by individuals of an age-corresponding control group) were processed using the Russian version of LIWC with users’ dictionaries. Unlike current studies, in designing the model we mostly made use of features that are not significantly dependent on the content. This is because not all individuals who committed suicides are known to deal with the topic in their texts. The resulting model was shown to be 71.5% accurate, which is comparable with the state-of-the-art for English texts.

Keywords. Suicide language, internet texts, suicide predictors, text corpus, computational linguistics, Russian texts, RusPersonality.

1 Introduction

Over 800,000 people die of suicide every year. It is estimated that by the year 2020, this figure will have increased to 1.5 million [3]. It is considered to be one of the major causes of mortality during adolescence [34]. The data suggest that only 30% of suicidal individuals report their inclinations [34].

Thus there is a growing need for methods of identifying suicidal individuals. Speech analysis including quantitative analysis is known to be a valuable diagnostic tool.

In addressing this particular issue, suicide notes are most commonly used. Texts are automatically classified as suicide notes or texts of other genres [13; 16; 24; 27; 28]. There are attempts to design models to distinguish between genuine and fake texts of this genre using qualitative text parameters as features [12] as well as to identify the differences between suicidal notes by individuals who eventually committed

doi: 10.13053/CyS-21-2-2721

suicide and those who made a suicide attempt [10].

Despite the ultimate value of studying suicide notes, being rather short in length, they present little opportunity for a profound insight into the language of suicidal individuals. As important as these studies are for tackling both theoretical and practical issues, there is an obvious necessity to investigate various linguistic features of texts by suicidal individuals written throughout various periods of their lives in order to identify the predictors of suicidal behavior.

This could then be compared to the texts of a control group of individuals with a maximum similar education level, social status and other characteristics, but most importantly who did not commit suicide. Based on the data, mathematical models can be designed to allow one to predict the likelihood of an individual's suicidal behavior using the quantitative parameters of their texts as features. Studies identifying linguistic features of texts by individuals who committed suicide are commonly conducted using their literary texts [1, 15, 16, 23, 25, 32], less commonly – segments of speech of famous suicidal individuals [7]. However, there are certain restrictions associated with the nature of texts and their authors' personalities, which prevents the results from being extrapolated into the entire population.

Recently, given the growing use of the Internet and social media in particular, scientists have been able to access very valuable linguistic data containing texts by individuals who either committed suicide, attempted suicide, or articulated their suicidal contemplations. Text mining methods have been successfully employed of late to identify such texts as well as over the last decade in sociology in particular [22] (note that there have been attempts using text analysis to identify not only Internet users with suicidal tendencies but also those displaying other forms of destructive behavior, see [33] for details).

Machine learning has been used to identify Twitter texts dealing with suicide [4] and stressed individuals based on their tweets [11], etc. (see Review in [8]).

Most research has been performed for English texts except very few [5, 9, 17].

Note that there have been very few studies of internet texts by individuals who died of suicide and they are mostly case studies (see [14] for Review). Even though most studies have sought to identify internet texts by individuals who had contemplated suicide, not all of them are known to be reported [28], and hence in designing models to identify individuals with high risk of suicidal behavior it is important to focus particularly on implicit suicide markers, i.e. those parameters of texts (features) that are not significantly dependent on the topic and are thus known to be a valuable source of personal data of their authors [31].

The objective of this paper is to design a mathematical model which would enable the classification of texts by suicidal individuals (those with high risks of suicidal behavior) and by non-suicidal ones based on numerical values of selected linguistic parameters (features). Blogs by young people who died as a result of suicide and samples of texts by written speech of young individuals were employed.

Thus, our study aims to contribute to the literature on suicide prevention in social media by attempting to classify texts as those by suicidal individuals based on text parameters (features) that are not significantly dependent on the content. The novelty of the paper is that it is performed on Russian texts and no research of this kind is known of as part of suicide prevention in social media [18].

2 State of the Art

As was previously stated, linguistic analysis of the speech of individuals who committed suicide originally utilized authors' suicide notes [24]. Quantitative parameters of texts used to be manually calculated and later on, automatic text processing tools started being employed for speech analysis. One of the most common tools used for investigating speech of individuals who died of suicide is Linguistic Inquiry and Word Count (LIWC) developed by the American psychologist J. Pennebaker and colleagues [25]. LIWC calculates the proportions of certain grammatical, lexical, and semantic markers, as well as markers belonging to other categories (up to 80 text features depending on the version).

In earlier studies suicidal notes were mainly investigated. Thus using LIWC software, it was found that notes from completed suicides had fewer metaphysical references, more future tense verbs, more social references (to others) and more positive emotions than did the notes from attempted suicides [10], which seems to be caused by individuals feeling relieved after making a decision to take their own lives.

Collecting texts from other genres written by individuals who committed suicide is a rather daunting task as a lot of research has been conducted using literary texts by suicidal writers. E.g., Stirman & Pennebaker [32], using the LIWC software, found that in poetic texts by suicidal individuals, written at different periods of time, the pronoun "I" is more frequently used compared to the texts by the test group. As time went by, suicidal individuals were seen to use fewer "we" pronouns as well as interaction verbs (e.g., talk, share, listen), but contrary to the common belief there were no statistically significant differences between the suicidal individuals and the test group in the number of words describing negative emotions. It is argued that the results are consistent with the suicide genesis theory, which postulates a connection between suicidal behavior and a growing alienation from other people.

In [1] it was shown that as an individual is just about to commit suicide there is an increase in first-person singular pronoun use, a decrease in first-person plural pronoun use, and an increase in negative emotion word use. In contrast, in [16] there was an increase in words expressing positive emotions, while words associated with causation and insight became less frequent.

Using LIWC, it was also found that suicidal individuals used more abstract words, fewer words overall, more verbs, fewer words relating to "Death", increased number of negations *(not, no)* and so on (see [18] for review).

Apart from literary texts, diaries by these individuals are employed in analysis [14]. Beside the descriptive method, the approach has been employed recently which classifies texts as those by suicidal individuals/not by suicidal individuals (e.g., see [23]). Texts are labelled using NLP tools, and classified using machine learning methods. Hence in [23] the models with the accuracy of 70.6% were obtained using song lyrics by suicidal individuals who committed suicide and those who did not (type/token ratio, proportions of first and second person pronouns, proportions of different vocabulary fields, etc. were employed as features).

The features that displayed differences between texts by suicidal and non-suicidal individuals were numbers (most useful across numerous algorithms), n-grams, the first-person singular + mental verb, concrete nouns, neutral terms, sensual words, total polar value semantic class features, and the first-person singular and passive construction syntactic features. The authors concluded that apart from expanding the corpus, "it would perhaps be fruitful to extend the analysis to other types of features and new lexicons since it has been demonstrated that this task could be solved using NLP" [23, p. 684].

Although written texts are mainly analyzed, there are individual studies which evaluate transcripts of recorded oral speech. A 2016 study [26] also argued the importance of using NLP techniques in automatic classification of texts, where the authors, using semi-supervised machine learning methods, recorded and analyzed the conversations of 30 suicidal adolescents and 30 matched controls.

The results shown that the machines accurately distinguished between suicidal and nonsuicidal teenagers. In another 2016 study of narrative notes, it was found that "incorporating a simple natural language processing strategy improved the ability to estimate risk for suicide and accidental death" [21].

There have recently been studies dealing with the analysis of internet texts by individuals with suicidal tendencies (e.g., see [2]). Publicly accessible blogs or Facebook provide new data sources for the study of suicidal behavior [17], however computational methods have only been used in a small number of suicide communication studies [4]. As was noted above, effort has been made to identify suicide-related content [4]. Some papers analyzed blog posts of individuals who died as a result of suicide (mostly case studies).

In [17] 193 blog entries of a 13-year-old boy posted during the year prior to his suicide were analyzed using the Chinese version of LIWC (CLIWC). Some language patterns related to suicide are similar in both Chinese and English, such as the use of first-person singular pronouns. Progressive self-referencing appeared to be a primary predictive of suicide.

Thus, in studies of texts by suicidal individuals, there is a clear shift from descriptive papers, which document differences in texts by suicidal individuals and control groups, to text classification studies of texts written by suicidal/non-suicidal individuals which make use of various machine learning algorithms. Though earlier papers analyzed mainly literary texts, recently social media texts have become a major real consideration.

Features were largely those extracted using LIWC, but some other papers employ well-known text representations like standard bag of words, character/word n-grams, etc. The overall consensus is that such techniques are fairly suited to furthering the development of methods to identify individual's suicidal tendencies based on analyzing their texts.

As social media takeover our lives, it is essential that there are methods in place to identify suicidal tendencies using Internet based texts. Unfortunately, no studies are known of that classify texts of individuals into suicidal or non-suicidal categories using Internet based texts from individuals who committed suicide and from a control group. In our study we are seeking to change this.

Additionally, most suicide studies have focused on English materials, while there is a clear lack of research into suicidal individuals of other cultural backgrounds [17].

3 Materials and Methods

3.1 Materials

In order to address this issue, what we basically need is a corpus of texts by suicidal individuals and those from a control group [18].

The publicly available blog entries of a suicide case were used as data for this study. The corpus of Russian texts *RusSuiCorpus,* written by individuals who committed suicide, currently contains texts by 45 Russian individuals aged from 14 to 25. The total volume of the corpus is 200 000 words. All the texts are manually collected and represent blog posts by individuals who committed suicide (blogs from *LiveJournal*). The fact that suicides were actually committed was checked by analyzing friends' comments, media texts, etc. Blogging is a prevalent form of communication in expressing emotion and sharing information, therefore it was chosen.

Being as the texts contained different numbers of words from each author, all the texts of blogs were joined into a single text with the length of about 200,000 words. The resulting text was divided into segments of 200 words making a total of 1,000 texts.

Ethics approval was obtained from the Human Research Ethics Committee for Non-clinical Faculties at Voronezh State University, Russia.

Samples of natural written speech from 1,000 students of various Russian universities, all a part of RusPersonality [19] (the first corpus of Russian texts containing wide metadata with the authors' personal information), were used. The average text length was 200 words with a total word count of 198045 words (further on NSUIC). All the texts were stream-of-consciousness essays. The authors were instructed to write whatever comes to their minds first and do in the manner they would normally do to their friends on social media.

We deliberately avoided using blogs for comparison because our goal in this study was to compare samples of natural written speech from individuals who died as a result of suicide with those from individuals whose results of psychological testing, identifying auto-aggressive behavior (with suicide as its extreme), are known [20]. It is obviously challenging to obtain this information about bloggers. At this point of the analysis we deliberately chose to include texts displaying varying risks of auto-aggressive behavior into the NSUIC corpus.

In addition, even though they were a part of different genres, the compared texts are quite homogeneous: they are all samples of what is called natural written speech, which is generally unrehearsed (the control group was instructed to write whatever first comes to mind for 30 minutes). Before proceeding to designing the model, the text corpora were divided into a learning set (900 texts in NSUIC, 900 in SUIC) and test set (100 in SUIC and 100 in NSUIC) texts.

3.2 Text Processing

All the texts were labelled using LIWC 2007 with the Russian users' dictionary. Users' dictionaries were also compiled. We selected features that were not significantly dependent on the topic of a text on purpose.

Hence the following features were selected: general LIWC descriptor categories (words per sentence (WPS), percent of words longer than six letters (Sixltr), from dimension I STANDARD LINGUISTIC DIMENSIONS (total function words, total pronouns, adverbs, prepositions, etc.), II PSYCHOLOGICAL PROCESSES (Social processes with subcategories, affective processes with subcategories, cognitive processes with subcategories, perceptual processes with subcategories and so on), AllPunc (the proportion of all the punctuation marks in a text overall and each mark individually).

Users' dictionaries were also compiled according to the user manual:

- *Deictic,* a dictionary of demonstrative pronouns and adverbs (where 1 feature represents the proportion of words per the total word length of a text), a dictionary of intensifiers and downtowners
- *Intens* (2 features), a dictionary of perception vocabulary
- *PerceptLex* (1 feature), a dictionary of pronouns and adverbs describing the speaker - *Ego* (I, my, in my opinion; 1 feature), and a dictionary of emotional words
- *Emo* (negative and positive; 2 features). All in all, there are 85 features.

The users' dictionaries were compiled using available dictionaries and Russian thesauri. As a Russian dictionary that came with the software was a translation of a corresponding English dictionary and did not stand independent testing, there are doubts as to the semantic category of the second group and thus they have to be evaluated independently and objectively, so in turn we have to check it manually and make some corrections.

4 Results and Discussion

4.1 Mathematical Processing of the Results of Text Analysis

Mathematical processing of the results of text analysis was performed using the professional software SPSS 13.0.

Originally features with a frequency of less than 50% were excluded from the total list (in the SUIC as well as in the NSUIC corpora). The frequency of a text parameter (feature) is the percentage proportion of the number of non-zero values of a certain feature compared to the total number of the analyzed texts in the corpus the feature was computed for.

A preliminary data analysis of the SUIC and NSUIC corpora, using the Shapiro-Wilk test, showed that most of the features have a non-normal distribution. In order to identify statistically significant differences between the features from SUIC and identical features from NSUIC, a method of comparing dispersions of the analyzed sets was employed. For that a one-factor dispersion analysis was used (ANOVA).

Originally we made use of Kruskal-Wallis one way analysis of variance on ranks that is a non-parametric alternative to F-criterion in our one-factor dispersion analysis. In the Kruskal-Wallis criterion, average ranks of each of the groups are compared with the total rank computed using all of the data. The Kruskal-Wallis test for the significance level $p<0.05$ allowed us to identify differences in the medial values of the groups of the features from SUIC and NSUIC corpora.

After a significant difference between the overall groups has been identified, it is then advisable to compare the average values of the existing groups. This is called a posteriori criterion of pairwise comparison [30].

Table 1. Features selected for designing the model and their calculation values

Features /Values of tests	**Function words**	**Pronouns**	**Verbs**	**Preposition**	**Conjunctions**
Frequency, % SUIC\ NSUIC	99.9\100	99.6\99.71	99.8\99.81	99.31\99.81	99.8\99.33
Kruskal Wallis: H p-level	82.232 <0.001	18.886 <0.001	102.932 <0.001	72.381 <0.001	258.808 <0.001
Median: SUIC\ NSUIC	**49.500** 47.075	12.000 **13.010**	**15.000** 12.900	12.000 **12.745**	**12.000** 9.480
Tukey: q p-level	12.824 <0.001	6.145 <0.001	14.346 <0.001	12.029 <0.001	22.749 <0.001
Fit criterion of Kolmogorov-Smirnov SUIC\ NSUIC	0.033\0.057	0.057\0.018	0.055\0.03	0.056\0.031	0.062\ 0.018
Features	**Cognitive processes**	**Inclusive**	**Comparison**	**Space**	**Comma**
Frequency % SUIC\ NSUIC	99.8\99.9	99.5\98.66	99.7\100	99.5\99.71	99.7\99.33
Kruskal Wallis: H p-level	123.480 <0.001	188.843 <0.001	35.480 <0.001	18.994 <0.001	75.673 <0.001
Median: SUIC\ NSUIC	**21.500** 19.345	**8.500** 6.640	19.000 **20.085**	10.000 **10.695**	**12.000** 10.530
Tukey: q p-level	15.713 <0.001	19.431 <0.001	8.423 <0.001	6.162 <0.001	12.301 <0.001
Fit criterion of Kolmogorov-Smirnov SUIC\ NSUIC	0.041\0.024	0.065\0.02	0.04\0.034	0.06\0.03	0.066\ 0.027

For that the Tukey test with p < 0.05 was used, which is a modified Student criterion.

As a result, based on the data of two tests (Kruskal-Wallis and Tukey), we excluded the features that did not meet neither the Kruskal-Wallis nor Tukey criterion from those features which passed the original analysis, and thus whose values do not have statistically significant differences in the SUIC and NSUIC corpora.

In the next phase, as part of the mathematical analysis of the data, we checked whether the distributions of the features selected at the first two stages from SUIC and NSUIC corpora are normal. One of the most effective criteria for testing the normality of the distribution is the Kolmogorov-Smirnov test that is more efficient than the alternative criteria and is designed for large selections [30].

Using the standard procedures of the SPSS software as well as a visual analysis of data by designing distribution histograms of the features selected by this point of the analysis from both corpora, we determined whether the empirical distributions of the analyzed features were normal. For this we computed a fit criterion of Kolmogorov-Smirnov / Lillefors and compared it with the critical value typical of a data set of this size. The critical coefficient calculated according to [30] in the Kolmogorov-Smirnov test for a data set with a dimensionality of n = 1000 and significance level p < 0.01 is Dc = 0.033.

4.2 Features Selected for the Model

The calculation results indicated that almost all of the features that are normally distributed in NSUIC, are not in SUIC. A visual test of the distribution of such features using histograms showed that the experimental histograms are asymmetric. In order to account for the differences in the type of distribution of the same features in SUIC (S) and NSUIC (N) corpora, it was decided that the features with the fit criterion of Kolmogorov-Smirnov/Lillefors of no more than the critical value (Dc = 0.033) in one of the selections is excluded out of those features left following the first stage of the mathematical processing and those no more than the doubled value of the critical coefficient, i.e. 0.066 in the second selection.

Hence, in order to design a model to appropriately measure the likelihood of a text being written by an individual who committed suicide, only feature which met the following criteria were employed:

1. Highly frequent ones;
2. Those that passed the Kruskal-Wallis and Tukey tests;
3. Those that, according to the Kolmogorov-Smirnov/Lillefors criterion, have a distribution close to the normal one where p<0.01, considering the initial assumptions.

The features that were selected in accordance with all of the above criteria, including their descriptive statistics are identified in Table 1.

As the analysis indicated, out of 130 features only 10 met all of those requirements (see Table 1). We found that in texts by suicidal individuals, compared to those by individuals from a control group, there are more *function words, verbs, conjunctions, words describing cognition overall, inclusion words, more commas, fewer prepositions, more words describing comparison, words describing space* and *pronouns.*

It was previously shown [20] that texts produced by individuals with a greater likelihood of self-destructive behavior (suicide is its extreme) typically show less lexical diversity, fewer prepositions, more pronouns overall (and particularly personal ones), a higher coefficient of coherence (due to more conjunctions and deictic particles). Blogs also displayed fewer prepositions and more conjunctions in suicidal individuals, but also fewer pronouns. The difference between these values was also less significant than for the other features.

Let us assume that texts by suicidal individuals are more abstract, contextual, have fewer spatial references, which we think is indicative of their self-centeredness.

The following disparities, which were not included in the model due to these features failure of the normality test, are worth mentioning: texts by suicidal individuals have more negations, fewer words describing social and perceptive processes (particularly vision), more vocabulary from the LIWC group “Body”, fewer words for positive emotions, and more words for negative emotions.

All in all, the above data are consistent with the hypothesis that suicidal individuals are more self-centered and less focused on seeing the world around them. However, unlike a lot of studies of English texts, Russian texts were not found to contain more "I" pronouns compared to those in the control group. Thus our significant finding is not that there are more self-references, but instead that the authors are more self-centered and less focused on the world around them.

4.3. Designing the Model

Let us denote a set of elements (i=1.10) selected for designing the above model as S_{Si} and S_{Ni} (i=1..10). These elements are mean values of ten selected features of the text from SUIC and NSUIC corpora respectively.

We can safely say that S_S and S_N are a set of numerical values [S_{S1}… S_{S10}] and [S_{N1}… S_{N10}] whose elements are mean values of the features from SUIC and NSUIC corpora.

For a text under analysis, a set of values of the same 10 features S_T must be determined.

The deviation of a set S_T from a set of numerical values S_S and S_N that are typical of the texts from SUIC and NSUIC corpora respectively is determined as follows.

We calculate the deviation of a set of values $\mathbf{S_S}$ from a set of values $\mathbf{S_T}$ as follows:

$$\chi_S^2 = \frac{1}{n}\sum_i^n \frac{(S_{Si} - S_{Ti})^2}{S_{Si}}. \quad (1)$$

Similarly, let us determine the deviation of the distribution S_N from S_T:

$$\chi_N^2 = \frac{1}{n}\sum_i^n \frac{(S_{Ni} - S_{Ti})^2}{S_{Ni}}. \quad (2)$$

Let us assume that in order to determine which type (suicidal/non-suicidal) a particular individual is, it would suffice to χ_N^2 and χ_S^2. If ${\chi_N^2}/{\chi_S^2} > 1$ than the text under analysis would have been written by a suicidal individual.

4.4 Accuracy of the Model

Checking the model on the test selection showed that it was 71.5% accurate (according to the number of the correctly classified texts), with a baseline of 50%. Reported accuracy results are comparable with the ones obtained with English texts (70.6) [23].

5 Conclusion and Future Work

The suggested approach showed to be fairly accurate in classifying texts, even despite the fact that we selected the features maximally independent of the content (proportions of commas, function words, etc.), which indicates that natural language processing and data mining are promising for use in the identification of suicidal behavior. The proposed method certainly has some restrictions associated with the varying number of individuals in SUIC and NSUIC as well as a relatively small number of features.

There are plans to extend SUIC and the list of features using common tools for labelling texts, including those developed by the team of authors, as well as tools for syntactic labelling of texts. It is known that «the use of syntactic n-grams… gives good results when predicting personal traits» [29]. We are also looking into utilizing certain indicators of lexical diversity that, as shown in [31], are critical for identifying the psychological condition of authors in emotional auto-reflexive writing.

Additionally, there is more work that needs to be done to validate and adapt the Russian version of LIWC.

In the future we plan to compare the proposed model with well-known text representations like standard bag of words, character/word n-grams, as well as with different weighting schemes and different learning algorithms to appreciate and compare the real predictive potential of the proposed model. It would have been interesting to note how the 10-features model performs with respect to the full (130 features) model or with respect to models using standard feature selection techniques from the machine learning area [23, 28].

We also plan to perform a comparative analysis of texts by suicidal individuals with high and low risks of autoaggressive behavior using the corpus *RusPersonality* [19, 20].

In this corpus the negative and positive texts have a slightly different composition: the positive texts come from a reduced group of authors with a sparse range of ages; the negatives texts, however, come from a more controlled group with a bigger base of authors. As such there are plans to compare the blog posts of suicidal individuals with those written by the control group.

Since "one of the main tasks of computational linguistics is to provide models for the development of applied systems with various kinds of automatic linguistic analysis" [31], what we ultimately are seeking to do is to design software for assessing risks of suicidal behavior based on the linguistic parameters of texts, which would be instrumental in online analysis of internet texts and could potentially send letters of warning to authors of texts or to their family and friends on social media. This research could also be applicable in the psychological assessment of suicide risk.

Acknowledgments

This research is supported by the grant of the Russian President awarded to young Russian scientists (Candidates of Sciences), project № МК-4633.2016.6 "Predicting Suicidal Tendencies of Individuals Based on their Speech Production".

References

1. **Baddeley, J.L., Daniel, G.R., & Pennebaker, J.W. (2011).** How Henry Hellyer's use of language foretold his suicide. *Crisis*, Vol. 32, No. 5, pp. 288–292.
2. **Barak, A. & Miron, O. (2005).** Writing characteristics of suicidal people on the Internet: A psychological investigation of emerging social environments. *Suicide and Life-Threatening Behavior*, Vol. *35,* Num. 5, pp. 507–524. DOI: 10.1521/suli.2005.35.5.507.
3. **Bertolote, J.M. & Fleischmann, A. (2009).** A global perspective on the magnitude of suicide mortality. *Oxford Textbook of Suicidology and Suicide Prevention, A Global Perspective*, Oxford University Press, pp. 91–99.
4. **Burnap, P., Colombo, W., & Scourfield, J. (2015).** Machine Classification and Analysis of Suicide-Related Communication on Twitter. *Proceedings of the 26th ACM Conference on Hypertext and Social Media*, Guzelyurt, TRNC, Cyprus, pp. 75-84. DOI: 10.1145/2700171. 2791023.
5. **Desmet, B. & Veronique, H. (2014).** Recognizing Suicidal Messages in Dutch Social Media. *Proceedings of Ninth international conference on language resources and evaluation* (*LREC 2014*), Reykjavik, Iceland, pp. 830–835.
6. **Dethlefs, N. & Schoene, A.M. (2016).** Automatic Identification of Suicide Notes from Linguistic and Sentiment Features. *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* LaTeCH@ACL, Berlin, Germany, pp. 128–133.
7. **Fernández-Cabana, M., García-Caballero, A., Alves-Pérez, M., García-García, M., & Mateos, R. (2012).** Suicidal traits in Marilyn Monroe's fragments: An LIWC analysis. *Crisis*, Vol. 34, No. 2, pp. 124–130. DOI: 10.1027/0227-5910/a000183.
8. **Gomez, J.M. (2014).** Language technologies for suicide prevention in social media. *Proceedings of the Workshop on Natural Language Processing in the 5th Information Systems Research Working Days (JISIC 2014).* Quito, Ecuador, pp. 21–29.
9. **Guan, L., Hao, B., Cheng, Q., Yip, P.S., & Zhu, T. (2015).** Identifying Chinese Microblog Users with High Suicide Probability Using Internet-Based Profile and Linguistic Features: Classification Model. *JMIR Ment Health,* Vol. 2, Núm. 2:e17. DOI: 10.2196/mental.4227.
10. **Handelman, L. D. & Lester, D. (2007).** The Content of Suicide Notes from Attempters and Completers. *Crisis*, Vol. 28, No. 2, pp. 102–104. DOI: 10.1027/0227-5910.28.2.102.
11. **Homan, C., Johar, R., Liu, T., Lytle, M., Silenzio, V., & Ovesdotter-Alm, C. (2014).** Toward macro-insights for suicide prevention: Analyzing fine-grained distress at scale. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology,* Baltimore, Maryland USA, pp. 107–117.
12. **Jones, N. & Bennell, C. (2007).** The Development and Validation of Statistical Prediction Rules for Discriminating Between Genuine and Simulated Suicide Notes. *Archives of Suicide Research: official journal of the International Academy for Suicide Research,* Vol. 11, No. 2, pp. 219–225. DOI: 10.1080/13811110701250176.
13. **Leenaars, A.A. (1988).** *Suicide notes: Predictive clues and patterns*. Human Sciences Press, New York, USA.

14. **Lester, D. (2014).** *The "I" of the storm. Understanding suicidal mind*. De Gruyter, Warsaw, Poland.
15. **Lester, D. & McSwain, S. (2010).** Poems by a suicide: Sara Teasdale. *Psychological Reports,* Vol. 106, No. 3, pp. 811–812. DOI: 10.2466/ pr0.106.3.811-81.
16. **Lester, D. & McSwain, S. (2011).** A text analysis of the poems of Sylvia Plath. *Psychological Reports,* Vol. 109, No. 1, pp. 73–76. DOI: 10.2466/09.12.28.PR0.109.4.73-7.
17. **Li, T.M., Chau, M., Yip, P.S., & Wong, P.W. (2014).** Temporal and Computerized Psycholinguistic Analysis of the Blog of a Chinese Adolescent Suicide. *Crisis*, Vol. 35, No. 3, pp. 168–75. DOI: 10.1027/0227-5910/a000248.
18. **Litvinova, T. (2016).** Corpus studies of speech of individuals who committed suicides. *Russian Linguistic Bulletin*, Vol. 7, No. 3, pp. 133–136. DOI: 10.18454/RULB.7.16.
19. **Litvinova, T., Litvinlova, O., Zagorovskaya, O., Seredin, P., Sboev, A., & Romanchenko, O. (2016).** "RusPersonality": a Russian corpus for authorship profiling and deception detection. *Proceedings of International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT 2016)*, Saint Petersburg, Russia, pp. 1–7. DOI: 10.1109/FRUCT. 2016.7584767.
20. **Litvinova, T., Zagorovskaya, O., Litvinova, O., & Seredin, P. (2016).** Profiling a set of personality traits of a text's author: a corpus-based approach. *Lecture Notes in Computer Science,* Vol. 9811, pp. 555–562. DOI: 10.1007/978-3-319-43958-7_67.
21. **McCoy, Th.H., Castro, V.M., Roberson, A.M., Snapper, L.A., & Perlis, R.H. (2016).** Improving Prediction of Suicide and Accidental Death after Discharge from General Hospitals With Natural Language Processing. *JAMA Psychiatry,* Vol. 73, No. 10, pp. 1064–1071. DOI: 10.1001/j amapsychiatry.2016.2172.
22. **Montes y Gómez, M., López-López, A., & Gelbukh A. (1999).** Text mining as a social thermometer. *Proc. Text Mining workshop at 16th International Joint Conference on Artificial Intelligence (IJCAI'99),* Stockholm, Sweden, pp. 103–107.
23. **Mulholland, M. & Quinn, J. (2013).** Suicidal Tendencies: The Automatic Classification of Suicidal and Non-Suicidal Lyricists Using NLP. *Processing of International Joint Conference on Natural Language Processing*, Nagoya, Japan, pp. 680–684.
24. **Osgood, C.E. & Walker, E.G. (1959).** Motivation and language behavior: A content analysis of suicide notes. *The Journal of Abnormal and Social Psychology,* Vol. 59, No. 1, pp. 58–67. DOI: http://dx.doi.org/10.1037/ h0047078.
25. **Pennebaker, J.W., Chung, C.K., Ireland, M., Gonzales, A., & Booth, R.J. (2007).** *The development and psychometric properties of LIWC2007*. Texas, USA.
26. **Pestian, J.P., Grupp-Phelan, J., Bretonnel-Cohen, K., Meyers, G., Richey, L.A., Matykiewicz, P., & Sorter, M.T. (2016).** A Controlled Trial Using Natural Language Processing to Examine the Language of Suicidal Adolescents in the Emergency Department. *Suicide Life Threat Behav,* Vol. 46, pp. 154–159. DOI:10.1111/sltb.12180.
27. **Pestian, J.P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., Cohen, K.B., Hurdle, J., & Brew, Ch. (2012).** Sentiment analysis of suicide notes: A shared task. *Biomedical Informatics Insights,* Vol. 5, No. 1, pp. 3–16. DOI: 10.4137/BII.S9042.
28. **Pestian, J., Nasrallah, H., Matykiewicz, P., Bennett, A., & Leenaars, A. (2010).** Suicide note classification using natural language processing: a content analysis. *Biomed Inform Insights,* Vol. 3, pp. 19–28.
29. **Posadas-Durán, J.P., Markov, I., Gómez-Adorno, H., Sidorov, G., Batyrshin, I., Gelbukh, A., & Pichardo-Lagunas, O. (2015).** Syntactic N-grams as Features for the Author Profiling Task. *Notebook for PAN at CLEF 2015,* Toulouse, France.
30. **Salkind, N. (2007).** *Encyclopedia of Measurement and Statistics*. Sage Publications Inc., California, USA. DOI: 10.4135/9781412952644.
31. **Sidorov, G. & Castro-Sánchez, N.A. (2006).** Automatic Emotional Personality Description using Linguistic Data. *Research in Computing Science,* Vol. 20, pp. 89–94.
32. **Stirman, W. & Pennebaker, J. (2001).** Word use in the poetry of suicidal and non-suicidal poets. *Psychosomatic Medicine*, Vol. 63, No. 4, pp. 517–522.
33. **Villatoro-Tello, E., Juarez-Gonzalez, A., Escalante, H.J., Montes-y-Gomez, M., & Villaseñor, L. (2012).** A Two-step Approach for Effective Detection of Misbehaving Users in Chats. *Paper presented at the meeting of the CLEF (Online Working Notes/Labs/Workshop),* Rome, Italy.
34. **World Health Organization (2014).** *Preventing suicide: A global imperative.* WHO Publications, Luxemburg.

Article received on 14/11/2016; accepted on 17/03/2017. Corresponding author is Tatiana A. Litvinova.