Determination of auto-aggressive behavior using machine learning methods

Olga Kanishcheva[1,*,†], Nadiia Babkova[2,*,†], Dina Huliieva [2], Zoia Kochuieva[2,†] and Nataliia Ugolnikova[2,†]

[1] *Friedrich Schiller Universität Jena, Universitätshauptgebäude, Fürstengraben, 1, Jena, 07743, Germany*
[2] *National Technical University «Kharkiv Polytechnic Institute», 2, Kyrpychova str., Kharkiv, 61002, Ukraine*

Abstract

The article is devoted to researching the possibilities of using machine learning methods to detect auto-aggressive behavior in texts, in particular, based on data from Twitter. The paper analyzed various formal and informal signs using the "Suicidal Ideation on Twitter" dataset, during which the most significant for the identification of auto-aggressive behavior were singled out. Logistic Regression and Random Forest methods were used for classification, which demonstrated satisfactory results. Further research is planned, which will include the application of neural models such as CNN, RNN (LSTM), and BERT, to compare their performance with classical methods. The obtained results indicate the prospects of using machine learning methods to detect auto-aggressive behavior in English texts, which may be extended to the Ukrainian language in the future. The obtained results can be used to improve the quality of life and reduce social exclusion for persons with a tendency to auto-aggression.

Keywords

Textual Attribution, Auto-aggressive Behavior, Text Classification, Machine Learning

1. Introduction

The study of auto-aggressive behavior and its impact on the psyche of an individual is becoming an increasingly common problem in the modern world and requires a comprehensive approach to its understanding and solution. The increasing incidence of self-aggressive behavior, especially among young people and people with mental disorders, highlights the need for further research in this area. Methods of studying auto-aggressive behavior can include several approaches, such as clinical observations, psychological testing, questionnaires, and analysis of texts and language, which allows obtaining various data and determining factors that influence the occurrence of auto-aggressive.

CLW-2024: Computational Linguistics Workshop at 8th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2024), April 12–13, 2024, Lviv, Ukraine
* Corresponding author.
† These authors contributed equally.
kanichshevaolga@gmail.com (O. Kanishcheva); Nadjenna@gmail.com (N. Babkova); dgulieva@ukr.net (D. Huliieva); kochueva@kochuev.com (Z. Kochuieva); natalie.uns@gmail.com (N. Ugolnikova)
0000-0002-9035-1765 (O. Kanishcheva); 0000-0002-2200-7794 (N. Babkova); 0000-0001-8310-745X (D. Huliieva); 0000-0002-4300-3370 (Z. Kochuieva); 0000-0003-2322-0922 (N. Ugolnikova)

The exploration of psychological aspects of the interaction between language and speech, which in the most general form are usually divided into processes of speech production and perception, language understanding, and its acquisition. Analysis of the correlation between processes of generation and perception of texts, and their congruence with the mental and psychophysiological state of people involved in communication processes makes it possible to:

- establish whether the text belongs to one person or another;
- identify the personal characteristics of the author;
- determine his emotional state, inner position, and attitudes.

Qualitative research methods such as interviews and focus groups with individuals who have experienced auto-aggressive behavior can help to understand their motivations, experiences, and needs, which contributes to increased empathy and understanding of the problem. Combined methods that combine clinical and psychometric approaches with text and language analysis can provide a more comprehensive understanding of auto-aggressive behavior and contribute to the development of effective treatment and prevention strategies. The research actively uses modern technologies and tools for data collection and analysis, such as programs for computer analysis of text and language, which allows to automate the research process and obtaining more objective results. The analysis of texts and speech with the help of software allows to reveal of the emotional tone and semantic nuances, which helps to identify the risks of auto-aggressive behavior and develop individualized approaches to its treatment. Innovative research methods, such as functional magnetic resonance imaging (fMRI) and electroencephalography (EEG), make it possible to study the neurophysiological features of auto-aggressive behavior and find out its impact on the brain. The development of research methods of auto-aggressive behavior is important for understanding its causes, mechanisms, and impact on the mental health of an individual and contributes to the development of effective psychological and medical strategies for the treatment and prevention of this phenomenon.

2. Related works

Auto-aggressive behavior is a complex and multifaceted phenomenon that can be interpreted from different perspectives. Currently, there is an increase in the number of depressions accompanied by phenomena of auto-aggression. Auto-aggressive behavior, which includes all types of intentional self-harm, including suicide attempts, is a significant suicidal risk factor. Many authors consider non-suicidal auto-aggression as a predictor of subsequent suicidal behavior, which makes the study of various aspects of self-harmful behavior an urgent interdisciplinary task.

Self-injurious behavior is associated with both intrapersonal and interpersonal conflicts, and in both cases, researchers confirm the regulatory aspect of such actions. Self-harm can provide a sense of control by changing anxious or suicidal thoughts, or stopping dissociative experiences. In some cases, acts of self-harm are used as forms of self-directed anger and punishment. In addition, self-injurious behavior may serve other functions: influencing

others, seeking attention, or physically expressing emotional distress. The best-known model of self-harm is the affect or emotional regulation model, where the role of self-harm is to alleviate acute negative states such as depression, anger, and anxiety. This model is supported by experimental evidence indicating that: youth and adolescents who self-harm tend to report higher levels of negative affect than those who do not self-harm; acute negative affect usually precedes an act of self-injurious behavior; many report decreased negative affect and relief after self-harm; Most adolescents and young adults who self-harm report a persistent desire to relieve negative emotions as well as persistent difficulty regulating other emotions

Let's consider several approaches to the interpretation of this phenomenon:

1. Psychological approach: Auto-aggressive behavior can be considered from a psychological point of view, where the inner world and emotional state of the individual are studied [1].
2. Social approach: The perception of the surrounding environment and the influence of social factors are also important for understanding auto-aggressive behavior [2].
3. Neurobiological approach: The study of brain processes and chemical reactions can shed light on the biological aspects of auto-aggressive behavior [3].
4. Cognitive approach: Studying the role of thinking and cognitive processes can help in understanding and predicting auto-aggressive behavior [4].
5. Trauma theory: Some researchers consider auto-aggressive behavior as the result of mental trauma or stressful situations [5].

These approaches complement each other, creating a comprehensive understanding of auto-aggressive behavior from different perspectives. Self-aggressive behavior is a serious problem defined by personal actions aimed at self-harm and may include suicide. This problem is most often associated with mental disorders, stress, or psychosocial difficulties. Individuals who exhibit auto-aggressive behavior may withdraw from their normal social context and feel separated or unaccepted, which reinforces their need for self-destruction. Factors influencing auto-aggression include psychological problems, including depression, anxiety, and personality disorders. The duration and intensity of auto-aggressive actions may depend on the degree of internal strife, which may be caused by experienced stress or trauma.

The psychological consequences of auto-aggressive behavior can include deterioration of the emotional state, the resulting feeling of own unfitness, and exacerbation of mental disorders. Self-harm can become a means of expressing inner psychic strife when other methods of communication become insufficient. Auto-aggressive behavior can be impulsive and emotionally determined, and not always manifests itself in the form of an expression of intended suicide. Individuals with auto-aggressive behavior may experience internal conflicts and a sense of lack of control over their own lives. Auto-aggression can be an individual's attempt to control his internal stress or to find a way out of a difficult emotional situation.

For many individuals, self-aggressive behavior is a way of diverting attention from mental pain and stress. A large part of individuals with a tendency to auto-aggression may

face social isolation and a feeling of rejection in society. Support is important in the treatment of self-aggressive behavior and the ability to self-help can greatly facilitate the recovery process. Underestimating and ignoring the problem of auto-aggressive behavior can lead to serious consequences for mental health and threaten the life of the individual.

The main aspects of auto-aggressive behavior cover a wide range of manifestations and factors that can influence its development. Below is a description of typical forms of auto-aggressive behavior and factors that can contribute to its occurrence:

1. Physical manifestations: Auto-aggressive behavior can take physical forms, such as self-hitting, cutting the skin, attempts at self-harm.
2. Use of objects: A person may use objects intentionally to cause injury or bodily harm.
3. Self-traumatization: Self-hitting, hitting something hard, or bumping into dangerous objects are forms of self-traumatization.
4. Suicidal attempts: Some forms of auto-aggressive behavior can be associated with suicidal attempts and attempts to inflict fatal injuries on oneself.
5. Emotional self-destruction: Psychological aspects include emotional self-destruction, such as attempts to undermine one's own mental stability through self-insurance or other methods.

Factors that can influence the development of auto-aggressive behavior:

1. Mental disorders: Mental disorders such as depression, anxiety, and post-traumatic stress disorder can play an important role [6].
2. Social factors: Social independence, and the feeling of lack of support in important relationships can affect the risk of auto-aggressive behavior [7].
3. Injuries and stress: The impact of traumatic events, especially in childhood, can be a significant factor in the emergence of auto-aggressive tendencies [8].
4. Substance abuse: The use of alcohol or drugs can increase the risk of self-harm [9].
5. Genetic factors: Heredity may play a role in susceptibility to mental disorders that may influence auto-aggressive behavior [10].

3. Description of dataset

The Suicidal ideation on Twitter dataset [11] was used in the study. This is a dataset of tweets compiled based on speech material of two youth social groups: people with auto-aggressive accentuation and those without it (i.e., people belonging to the norm group). This data set is the first in which the main attention is paid to the division of suicidal thoughts into active and passive. These two types of suicidal thoughts require different treatment by specialists in the future, and therefore they should not be combined. This set of data is also the first in which passive suicidal thoughts are highlighted, which are no less dangerous than active suicidal thoughts and differ from active statements about suicide in the lexical content. Moreover, this data set is also the first to distinguish between sarcasm and suicidal ideation. Sarcasm is very common on Twitter, so this distinction is all the more necessary. The dataset consists of 81,519 tweets, of which 5,051 are marked as active suicidal thoughts, 5,055 as passive suicidal thoughts, 5,009 as sarcasm, 5,005 tweets related to

suicide, and 61,333 tweets not related to auto-aggressive behavior. The dataset consists of five classes (active suicidal thoughts; passive suicidal thoughts; sarcasm about suicidal thoughts; suicide-related tweets (awareness, news, suicide talk; other), and each class was randomized and then divided into three parts: 70 percent for the training set (57076 tweets), 15 percent for the validation set (1221 tweets), and 15 percent for the test set (12222 tweets).

For the implementation of experiments on the definition of auto-aggression, all texts included in the dataset were pre-processed. It is worth noting that preprocessing varied depending on the type of implemented classification models, but in general, it consisted of four stages. The first stage of processing included the tokenization of texts, in this case, the division into tokens took place by words. During the following stages of processing the texts of the dataset, normalization, and removal of punctuation marks and stop words were carried out, and the final stage of processing was the segmentation of the texts.

4. Data analysis and determination of the characteristics of auto-aggression

This section analyzes the texts of people with and without auto-aggressive accentuation. This analysis was carried out with the aim of identifying formal and informal signs characteristic of people with a tendency to auto-aggression. During the analysis, a comparison of speech characteristics of two social groups was made to identify fundamental differences in frequency elements.

As distinguishing speech signs of psychotypes, it is possible to consider units of language levels: phonemic, morphemic, lexical, and syntactic. However, within the framework of the study of written texts, it is rational to stop at the last three. The verbal material of the study was subjected to automatic processing and further interpretation of the obtained data by methods of linguistic analysis.

As a result of the analysis, it was determined that at the morpheme level it is possible to consider some word-formation models that prevail in the written speech of people with auto-aggressive accentuation, namely the use of the prefix anti- in the sense of "opposite" (**anti** – *experience with antidepressants best and worst i was put on paxil back in 2005 it was absolutely horrible a lot of side effects and very addictive it took me almost a year to ween myself off of it and there are even more side effects trying to get it out of your systemi m now on trintellix been using it for almost 3 months and so far it has been a good experience no major side effects besides a little nausea at firstmy wife had a terrible time on zoloft it seemed to make her postpartum depression worse and she would have really bad episodes of anxietywhat have you all tried and what has worked or didn t work for you*); prefix in- in the meaning "absence of something" in the case when words without these prefixes have a positive connotation (ineffective); negative prefix not- in case words without this prefix have a positive connotation (not easy).

In **Figure 1** presents the results of a partial analysis of texts containing auto-aggression. The diagram (**Figure 1**) shows that people with auto-aggressive accentuation are characterized by the use of a smaller number of prepositions, and a larger number of pronouns with a higher index of logical connection, which is achieved due to the use of a

larger number of conjunctions and deictic particles. It was also noted that compared to the group of norms, there are more verbs in the speech of suicides, among which a significant part is made up of non-factual ones. Another feature is a large number of adjectives, the use of which is twice as high in the group with auto-aggressive accentuation. Such a tendency is connected with the desire to verbalize mental experiences and emotions experienced by people of this social group.

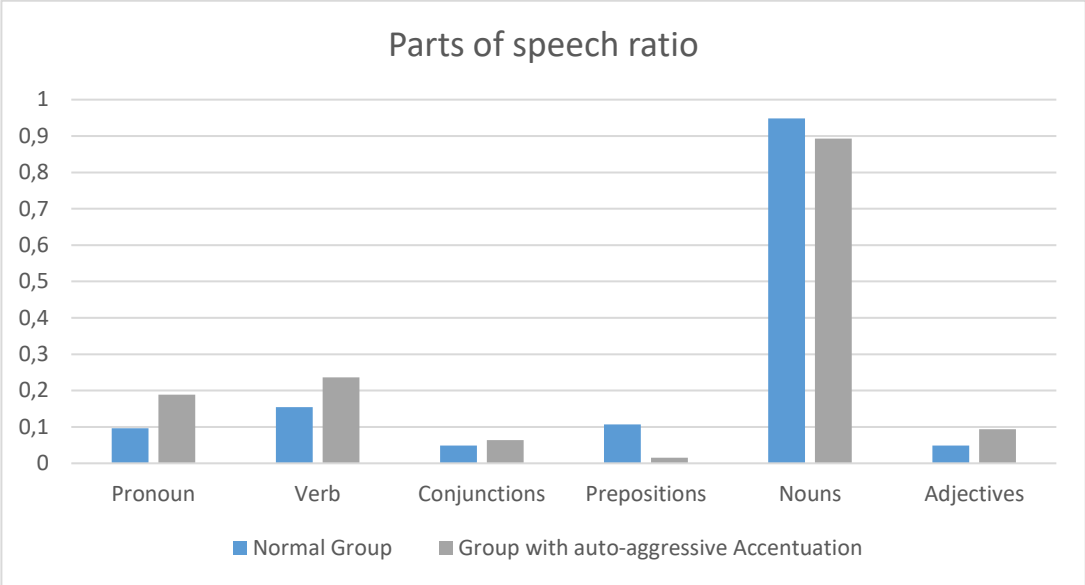

Figure 1: Ratio of parts of speech in the norm group and the group with auto-aggressive

It is also necessary to emphasize the predominance of I-group pronouns (**I, me, my** – *i wonder when i will do it i know its going to be the way i go out in my life i keep coming back to my depression i have periods in my life where i am happy but i always come back to this its exhausting and i just want to end it but i hate that my family will be pained by it i hate living here i hate existing*), which emphasizes the egocentricity of the written speech of the social group being studied.

At the lexical level, for texts of this social group, one can note the use of abstract vocabulary (**love** – *feeling overwhelmed and wanting to die i tried to kill myself today all i wanted was my boyfriend to help me somehow since he sees me on this depression spiral but all he says is it will get better i know it will but currently i feel trapped in my own home i have no friends and i constantly feel weak or unmotivated i have no desire to have sex with my partner and tend to find the worst in him lately i love him but some things he does amplifies my depression i dont want to die but when i get hurt lonely or overwhelmed the one thought i cant shake is killing myself,* **fear** – *i cant go on living i have been struggling with depression since junior high at 40 i understood a lot about how this came to happen and i am so done with being alive i am not sad anymore not angry not afraid but i have a big painful problem i am a father this makes it worse a living hell literally for me the only reason i am alive at the moment is the fear of hurting my kids an irrational fear because they have a good mom and life insurance can most likely help but i have something in the back of my head that tells me that my death would be a loss of experience*

and protection to them on the other hand i feel that i have no control over the universe and that me staying is pointless anyways i just want to have enough assurance that my presence wont be necessary anymore because is fucking hell and it hurts everywhere being alive, **pain** *– goodbye world this is it i cant deal any longer with my pain i am tired i was blessed with nothing in my life i am lost and stopped eating i am finally at peace with my decision i know no one cares and thats ok it just makes it that much easier,* **happiness** *– nothings working i recently got into a new hobby that i am very interested in and i ve been able to hang out with close friends every time we get to i have a lot of fun my boyfriend and i go to different universities so we dont get to see each other often but hes been starting to visit once a week ifi amlucky all these things contribute to my happiness but its been small temporary spurts of it other days i will have thoughts of self harm or jumping in front of a car things like no one will notice me gone or if i disappear it may be easier on my family since its one less mouth to feed one less tuition to pay for or if i disappear it wont make a difference in anyones life i just have a constant voice in the back of my head that tells me i dont do good enoughi ama nuisance to others and i should kill myself ironically i am very afraid of death so i go to self harm instead its like a chickens way out i try not to do anything rough enough that it will leave marks but sometimes bruises or small scars are left; i wonder when i will do it i know its going to be the way i go out in my life i keep coming back to my depression i have periods in my life where i am happy but i always come back to this its exhausting and i just want to end it but i hate that my family will be pained by it i hate living here i hate existing,* **egoism** *– i have nothing left to live for my narcissistic family hate me and my soul mate left me without reason i have no job no friends no family and i cant see any way through i simply just give up i want this post to show i love you scott i love you thom i love you tom and i love you camilla and i am sorry for the selfishness of what i am doing but there is no future for me goodbye*), with the help of which a person conveys his feelings, emotions and reasoning. In addition, it is worth paying attention to the large number of negations (*expressed using particles no and not*), as well as negatively colored lexemes (*selfish, disgusting, hate*) and the dominance of the author's negative assessment in relation to the objects discussed in the text (*He looks like a zombie*). This creates a tendency towards the predominance of texts with a negative tone among people of the social group being studied.

During the study, a sentiment analysis of the texts of the dataset used was carried out, where for each text the emotional coloring was determined in accordance with the following categories: neutral, positive and negative. On part of the dataset, a classifier was trained using semi-supervised learning to determine membership in a certain emotional category. After classification, the proportion of texts of each of the presented sentiment types in the dataset was calculated (**Figure 2**). It can be noted that the proportion of texts written by people with auto-aggressive accentuation and having a negative emotional connotation is 6 times higher than the proportion of the same texts, but already written by people in the norm group.

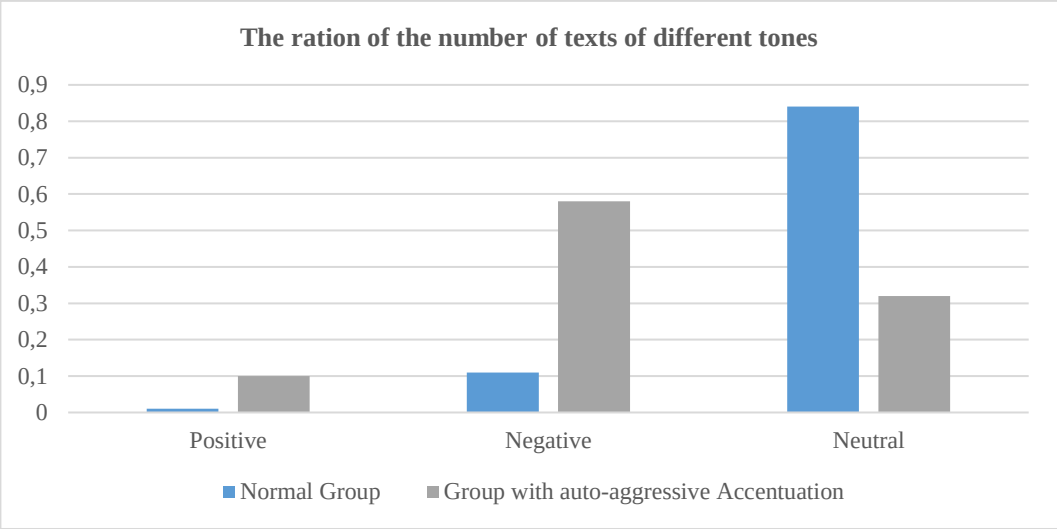

Figure 2: The ratio of the number of texts of different tones

In addition, people suffering from auto-aggression tend to use a limited set of lexemes in their speech, which indicates low lexical diversity. Another distinctive feature is the subjective attitude to reality (**focus on one's own inner world** – *i hate feelings at this point i think i am clinging onto life waiting for the right set of circumstances to just let go my feelings keep holding me back i hate them i hate having to feel i hate each and every one of them theres something comforting about the idea of nonexistence something comforting about the idea of nothing i long for it*), which is manifested in the use of predicates of attitude (**love** – *buying a gun tomorrow since stores are closed i ll spare you the details of my life since no one will give a shit anywayi amjust done with everything nothing helps yes i am well aware that i m not alone it doesnt make a shit of difference nothing can be done and i m sick of wasting life time money air as much as i love the people in my life they arent worth living with this shit nor are they of any help because they dont know what to do i tried knives but i cant bring myself to draw blood*, **hate** – *i can t do this anymore i have no hope i m a senior in college and i just want to die i don t fit in and i feel like no one cares about me everyday i wake up and i try to tell myself that today is a new day but it doesn t work i hate everything about myself and i feel so alone i have no one i don t care how i look and i don t even want to get up in the mornings i just want to end it all end all the suffering and the pain i just want to feel nothing for my anxiety and depression to leave me alone i don t know where to turn but i just can t keep doing this anymore*), feelings and internal state (**be afraid** – *i am so tired i just want to sleep but i m afraid of what i might dream of i guess its just my own fault i guess that all of my problems are my own fault ive heard thats supposed to make me feel better because it means i can work to get better and change things it only makes me feel worse because i know i wont people wont care if i die i ve asked some at least then i could rest i wouldnt be tired i wouldnt be*, **worry** – *i keep imagining killing myself i am a struggling student at risk of being retained this year i keep worrying about failing the exams and i even considered killing myself should i fail i kinda walked back on the latter thinking i ll get a job as a cook or something still thoughts of me slitting my neck jumping to my death overdosing on pills they stay and theyre inferring with my mind*, **sad** – *i have no idea the only reason i havent killed myself is because i dont want to*

make others sad i am not happy but i am not sad either i take other peoples idea of their own happiness and make it mine i feed off of other peoples idea of life because i dont have my own any relationship ive ever been in was only to give the other person what they wanted mei am so ready to give up because i truely honestly have no idea what i want out of life or what i want for my self i have no goals or ambitions or dreams of how i want my life to be i just do what others expect out of a person and that is simply to not give up on life, **bored** – *i am literally unable to stop thinking about it suicide i can not stop thinking about it ways to do it how it would feel the relief of finally having some blood pump through my veins right before i hit the groundi am so bored and unhappy with my life that the act of ending it would be the only thing to provide sufficient excitementi amdepressed and have been since 17 the fact is i dont feel anything at all except empty boredomi amon autopilot and want to die i dont even know if id consider myself unhappy happy and sad are emotions and i am completely void of emotionfound this subreddit after weeks of obsessing over iti think i will flip a coin*), and personal pronouns.

In the course of the work, it was found that the texts of people with auto-aggression are distinguished by the prevalence of sentences (their greater length) and a higher readability index, which indicates the construction of sentences that are more understandable for perception. Communication within a sentence is carried out primarily through conjunctions; the use of coordinating connections is more frequent than subordinating ones (getting into a partial program is impossible if you weren't just hospitalized guess I gotta go attempt again so much for trying to get better; what do I hold on for there is still hope).

You can also note some stylistic features that are associated with syntax, namely a violation of sentence construction. People with auto-aggressive accentuation tend to put in the initial position in the text components that are important to them, namely: the expression of feelings, emotions, self-awareness (*no one to talk to if there was one i would be too anxious and get a panic attack i can't even see therapists anymore my anxiety and depressions got so bad i cant talk to anyone anymore also i ruined my last friendship just now so i cancelled every therapist appointment i would have gone to just for my friend all motivation is gone and the suicide thoughts never were so damn reali amjust so hopeless crying all day in my dark appartment no friends family or job also my savings will be gone very soon and i end up on the streets its sad that death seems like the only logical option to me i am so lonely it has become physical pain i don't know what to do i have no one).*

After the analysis, a selection of characteristics was made that will be used for software implementation of diagnostic models. During the selection of characteristics, it was taken into account that not all linguistic features that indicate that a person belongs to an auto-aggressive type can be analyzed by computer methods and included in the diagnostic model. Therefore, the main part of the selected features are formal. Thus, the following formal and informal characteristics were selected for software implementation, describing the idiostyle of the generalized linguistic personality of the social group being studied:

1. morphological parameters:

- frequency of occurrence of various parts of speech (pronouns, in particular I-group pronouns; verbs; adjectives, etc.);
- the ratio of different parts of speech - Flesch-Kincaid readability index.

2. lexical parameters:

- degree of lexical diversity;
- the tone of the text;
- quantitative assessment of the frequency of occurrence of a word in context (TF-IDF).

3. syntax parameters:

- average length of sentences.

The study of the relationship between morphological, lexical, and syntactic characteristics and the target variable (the presence of auto-aggressive accentuation), as well as a correlation analysis of these parameters, are considered in the work of Pennebaker [12]. In **Figure 3** presents the parameters arranged in increasing order of the degree of correlation with the target variable. Note that one parameter — the frequency of occurrence of nouns — was excluded from the characteristics that were used to build diagnostic models due to the correlation coefficient being too low.

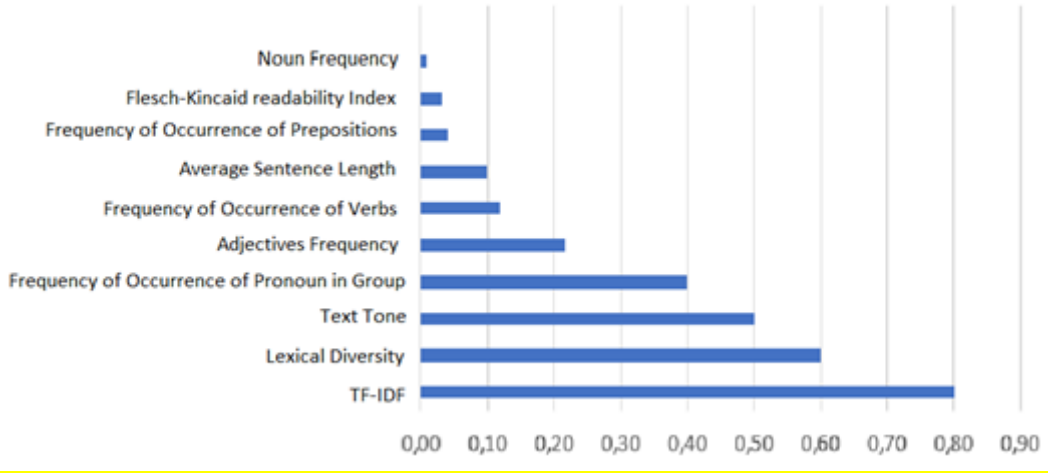

Figure 3: Correlation analysis between selected parameters and the target variable (presence of auto-aggressive accentuation)

Thus, the most informative in the context of determining a person's propensity for auto-aggressive behavior are such parameters as a quantitative assessment of the frequency of occurrence of a word in the context, lexical diversity, and tone of the text. Such results may indicate the key role of the lexical level in determining the personality type from the text. The choice of words, their frequency, and compatibility — all reflects the individual style and character of the author, and can also indicate his personality traits, mental processes, and emotional state. Thus, the TF-IDF metric used to assess the importance of a word in context, i.e. its rarity in the corpus and frequency in a particular text indicates the correlation of this accentuated psychotype with the author's vocabulary, the ratio of rare

and unique words for the corpus as a whole used by him in the text. The degree of lexical diversity reflects the author's ability to use different words and expressions to convey his thoughts and ideas, and the richness of his vocabulary. The tone of the text, in turn, can say a lot about the author, his attitude to the topic described in the text, his emotional state, and beliefs. Since speech is mainly dominated by the use of neutral, cross-style, commonly used vocabulary, against its background one can especially clearly note the characteristic stylistically colored words and expressions characteristic of one or another type of personality.

5. Experiments and analysis of the results obtained

Before training diagnostic models, texts were vectorized. In this work, the following vectorization methods were used:

1. presentation of the text in the form of a list of numbers, each of which is a numerical characteristic of one or another linguistic feature characteristic of the speech of people with auto-aggressive accentuation. Before creating the vector, the data was normalized, as well as the distribution of the degree of importance of each parameter based on the results of correlation analysis. The final pool of parameters for this type of vectorization consisted of the following characteristics: degree of lexical diversity; the tone of the text; frequency of occurrence of I-group pronouns; frequency of occurrence of pronouns, verbs, adjectives, prepositions; Flesch-Kincaid Readability Index and average sentence length.
2. vectorization of the text using the TF-IDF metric, which allows you to calculate the frequency of a word in the text, taking into account the degree of its "uniqueness" in the entire corpus. TF (Term Frequency) reflects the frequency of occurrence of a word in a document and is calculated as the ratio of the number of occurrences of a word to the total number of words in the document. Thus, TF shows how important a word is to a given document. IDF (Inverse Document Frequency) reflects the importance of a word for a collection of documents and is calculated as the logarithm of the ratio of the total number of documents to the number of documents containing a given word. Thus, IDF shows how unique a word is in the context of a collection of documents. The final TF-IDF value for each word in the document is calculated as the product of TF and IDF.

Next, the assembled corpus was balanced relative to the accentuation being diagnosed, which allows for more accurate results to be achieved. After this, the corpus was divided into a training and test set. The texts were divided randomly in a 4:1 ratio, with 80% of the entries assigned to the training set and 20% to the test set.

To identify the most effective automatic method for diagnosing auto-aggressive tendencies, two classification methods were selected from the author of the written text: Logistic Regression and Random Forest.

Models for diagnosing a personality's propensity for auto-aggressive behavior were trained on a training set and then tested on test texts, which made it possible to draw a conclusion about the performance of each of the created models.

When assessing and comparing the performance of the presented methods, metrics were used that are traditionally used in assessing the quality of classification: Precision (accuracy), Recall (completeness), and F-score (F-measure). The assessment was carried out for each class separately, and then the weighted average method was used to demonstrate overall performance (Table 1).

Table 1
Classification results

	TF-IDF			a vector consisting of the values of various linguistic features		
	Precision	Recall	F-score	Precision	Recall	F-score
Random Forest	0.87	0.84	0.84	0.95	0.84	0.89
Logistic Regression	0.96	0.95	0.95	0.89	0.83	0.86

Thus, as a result of the study, Logistic Regression copes best with the task of identifying auto-aggressive tendencies when vectorizing text using TF-IDF (F-measure = 0.95).

6. Conclusions and further research

In this work, an analysis of the task of identifying auto-aggressive behavior using additional methods of machine learning was carried out. Based on the "Suicidal Ideation on Twitter" dataset, formal and informal signs were analyzed and those most significant for identifying auto-aggressive behavior were identified. Several classification methods have shown results from Logistic Regression and Random Forest. We are planning to apply neural models such as CNN, RNN (LSTM), and BERT to this dataset, and compare the results with classical machine learning methods. In addition, this research has revealed the promise of these methods for identifying auto-aggressive behavior in English texts, and in the future, it is planned to scale it up for the Ukrainian language.

References

[1] Klonsky, E. D. "The functions of deliberate self-injury: A review of the evidence." Clinical Psychology Review 27.2 (2007): 226–239.

[2] O'Connor, R. C., J. Pirkis, and G. R. Cox. The International Handbook of Suicide Prevention. John Wiley & Sons, 2020.

[3] Mann, J. J., C. Waternaux, G. L. Haas, and K. M. Malone. "Toward a clinical model of suicidal behavior in psychiatric patients." American Journal of Psychiatry 156.2 (1999): 181–189.

[4] Rudd, M. D. "Fluid vulnerability theory: A cognitive approach to understanding the process of acute and chronic suicide risk." In Suicide science: Expanding the boundaries, 2006, pp. 275–296.
[5] van der Kolk, B. A., J. C. Perry, and J. L. Herman. "Childhood origins of self-destructive behavior." American Journal of Psychiatry 148.12 (1991): 1665–1671.
[6] Joiner, T. E., Jr. Why People Die by Suicide. Harvard University Press, 2005.
[7] Baumeister, R. F. "Suicide as escape from self." Psychological Review 97.1 (1990): 90–113.
[8] Nock, M. K., and M. J. Prinstein. "A functional approach to the assessment of self-mutilative behavior." Journal of Consulting and Clinical Psychology 72.5 (2004): 885–890.
[9] Wilcox, H. C., K. R. Conner, and E. D. Caine. "Association of alcohol and drug use disorders and completed suicide: An empirical review of cohort studies." Drug and Alcohol Dependence 76 (2004): S11–S19.
[10] Brent, D. A., N. M. Melhem, M. Oquendo, A. Burke, B. Birmaher, B. Stanley, ... and J. J. Mann. "Familial pathways to early-onset suicide attempt: Risk for suicidal behavior in offspring of mood-disordered suicide attempters." Archives of General Psychiatry 67.8 (2010): 801–810.
[11] Suicidal ideation on Twitter (dataset). URL: https://www.kaggle.com/datasets/natalialech/suicidal-ideation-on-twitter/data.
[12] Pennebaker, J.W., M.R. Mehl, and K. Niederhoffer. "Psychological aspects of natural language use: Our words, our selves." Annual Review of Psychology 54 (2003): 547–577.