УДК 004.83
DOI: https://doi.org/10.11603/mie.1996-1960.2020.1.11125

ЗАСТОСУВАННЯ МЕТОДІВ ОБРОБЛЕННЯ ПРИРОДНОЇ МОВИ ДЛЯ ВИЯВЛЕННЯ СИМПТОМІВ МЕНТАЛЬНОГО ЗАХВОРЮВАННЯ

С. Д. Погорілий, А. А. Крамов

Київський національний університет імені Тараса Шевченка

Здійснено порівняльний аналіз різних методів оброблення природної мови для виявлення симптомів ментального захворювання. Розглянуто принцип роботи та ефективність моделей оцінювання семантичної когерентності тексту (моделі тангенційності та некогерентності) для класифікації текстів здорових і хворих осіб. У роботі зазначається залежність точності моделей некогерентності та тангенційності від моделі семантичного представлення фрагментів тексту; підкреслюється недолік використання такої моделі в зв'язку з відсутністю можливості враховувати регулярне повторення фраз. Проаналізовано переваги та недоліки застосування комбінації моделей семантичного представлення елементів тексту для врахування постійних повторів його фрагментів. Обґрунтовано доцільність застосування лінгвістичних характеристик тексту пацієнта для підвищення точності класифікаторів виявлення симптомів захворювань та розрізнення їх типу. Розглянуто можливість аналізу частоти появи неоднозначних займенників у тексті для підвищення точності класифікації даних. Проаналізовано особливості застосування різних методів виявлення симптомів ментального захворювання для текстів англійською, німецькою та російською мовами. Запропоновано здійснювати оцінювання зв'язності тексту за допомогою графу узгодженості словосполучень. Здійснено експериментальну перевірку ефективності пропонованого підходу для побудови класифікаційної моделі порівняно з іншими характеристиками тексту.

Ключові слова: виявлення симптомів ментального захворювання, обробка природної мови, модель некогерентності, модель тангенційності, семантичне представлення елементів тексту, класифікаційна модель.

THE USAGE OF NATURAL LANGUAGE PROCESSING METHODS TO DETECT THE SYMPTOMS OF MENTAL ILLNESS

S. D. Pogorilyy, A. A. Kramov

Taras Shevchenko National University of Kyiv

Background. The process of the detection of the symptoms of mental illness is a complicated task that requires the appropriate level of the qualification of a specialist to solve it. One part of the diagnostics of such diseases is the analysis of the patient's speech. Alogia (the poverty of speech), the lack of the persistent focus on a topic, incoherent speech, permanent usage of metaphors can indicate the availability of appropriate symptoms. Thus, it is necessary to apply the different automated methods of the estimation of the patient's speech in order to detect some deviations from defined statistical data. Such methods fall into the category of natural language processing. Taking into account the lack of unified structure, the availability of ambiguous terms, the tasks of natural language processing cannot be solved with the usage of defined algorithms. Search for regularities and the detection of the connection between the text's elements are performed using the methods of machine learning: regression models, decision trees, deep learning (multilayer neural networks). Thus, it is advisable to consider state-of-the-art methods, based on different methods of machine learning, to detect the symptoms of mental illness by analyzing the patient's speech. The purpose of the work is the following: to perform the comparative analysis of different state-of-the-art methods of the detection of the symptoms of mental illness based on the methods of natural language processing; to make the experimental examination of the effectiveness of the proposed method based on the analysis of the connectivity of the text's elements.

Materials and methods. Results. According to the analysis of state-of-the-art methods, the semantic coherence is the main feature of a text to predict mental illness. Two different models based on the estimation of semantic coherence are considered: tangentiality model and incoherence model. The main idea of the tangentiality model consists in the detection of the persistent deviation of the topic of an answer from a question. A text is divided into windows — sets of words with a fixed length. Each window and question are represented as vectors using a pre-trained semantic embedding model — LSA. The similarity between a window and the question is calculated as the cosine distance between corresponding vectors. Using the set of calculated distances, linear regression is built. The steeper slope of a line indicates the deviation of the thoughts of a speaker from the whole topic of a conversation. In comparison to the tangentiality model, the incoherence model processes a text at the level of sentences. All sentences are represented as the average vector interpretation of its words; each word is represented as a vector using a pre-trained semantic embedding model. Then three different features are calculated to form a feature vector: minimum first-order coherence (minimum similarity between two sentences that is estimated as a cosine distance between corresponding vectors), maximum sentence length, and the frequency of the usage of additional uninformative words. This dataset is used to build a convex hull classifier that divides interviews of healthy and ill people. The key disadvantage of both mentioned models is the neglect of the repeats of phrases within a text. Moreover, such repeats can complicate the classification process. In order to solve it, the different combination of state-of-the-art semantic embedding models (Word2Vec, Sent2Vec, GloVe) with frequency algorithms (TF-IDF, SIF) can be used. The disadvantage of such an approach is the dependency on an additional corpus to calculate statistical data about the frequency of words' usage. As for the effectiveness of each model for different languages, it depends on the collected dataset and the unique features of a separate language. Except for the semantic coherence, other linguistic characteristics

can be taken into account to form a feature vector: linguistic complexity, linguistic density, syntactic complexity. Each of these characteristics can be represented with the corresponding set of metrics. Moreover, the frequent usage of ambiguous pronouns may also be taken into account because it can indicate the disorganization of the thoughts of a speaker.

The proposed method based on the graph of the consistency of phrases allows estimating the connectivity of a text — its cohesion. It takes into account the availability of coreferent objects and common terms within a text. The effectiveness of the suggested method was compared with other features of a text using pre-trained classification models. The results obtained can indicate that the proposed method may be used to calculate the connectivity feature for a model that predicts a mental illness.

Conclusions. As the main criteria to distinguish the texts of healthy and ill persons, the semantic coherence is used. The estimation of the semantic coherence is performed in the following models: tangentiality model and incoherence model. It is advisable to perform the semantic representation of the text's elements (sentences for the incoherence model and windows for the tangentiailty model) using the combination of different semantic embedding models with statistical algorithms (TF-IDF, SIF) in order to take into account permanent repeats of phrases. As for the effectiveness of the mentioned models for different languages, it depends on the semantic embedding model and the properties of a certain language.

In order to increase the accuracy of the classification model, other linguistic features should be taken into account: lexical density, lexical and syntactic complexity, connectivity. The method based on the graph of the consistency of phrases has been proposed to take into account the connectivity of a text. The experimental examination of the effectiveness of the proposed method in comparison with other features has been verified. The results obtained can indicate the expediency of the usage of the proposed method to increase the accuracy of a prediction model.

Key words: detection of the symptoms of mental illness, natural language processing, incoherence model, tangentiality model, semantic representation of the text's elements, classification model.

ПРИМЕНЕНИЕ МЕТОДОВ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА ДЛЯ ОПРЕДЕЛЕНИЯ СИМПТОМОВ МЕНТАЛЬНОГО ЗАБОЛЕВАНИЯ

С. Д. Погорелый, А. А. Крамов

Киевский национальный университет имени Тараса Шевченко

В работе осуществлен сравнительный анализ методов обработки естественного языка для определения симптомов ментального заболевания. Показано, что современные методы, основанные на алгоритмах обработки естественного языка, в качестве основного критерия прогнозирования заболеваний (шизофрения, биполярное расстройство) используют оценку когерентности текста. Под когерентностью текста подразумевается тематическая целостность его элементов, наличие постоянного фокуса вокруг темы доклада или диалога. Одним из критериев наличия когерентности текста является семантическая взаимосвязь фрагментов текста (фраз и предложений). Рассмотрен принцип работы и эффективность моделей оценки семантической когерентности текста (модели тангенциальности и некогерентности) для классификации текстов здоровых и больных лиц. Проанализировано возможное решение этой проблемы с использованием комбинации различных моделей семантического представления элементов текста, рассмотрены его преимущества и недостатки. Обоснована целесообразность использования лингвистических характеристик текста пациента (лексическое разнообразие, лексическая плотность) для увеличения точности классификатора определения симптомов ментальных заболеваний и различий их типа. Рассмотрена возможность анализа частоты появления неоднозначных местоимений в тексте для увеличения точности классификации данных. Проанализированы особенности применения различных методов определения симптомов ментального заболевания для текстов на английском, немецком и русском языках. Предложена оценка связности текста на основе графа согласованности словосочетаний, осуществлена экспериментальная проверка эффективности предложенного подхода по сравнению с другими характеристиками текста.

Ключевые слова: определение симптомов ментального заболевания, обработка естественного языка, модель некогерентности, модель тангенциальности, семантическое представление элементов текста, классификационная модель.

Вступ. За статистичними даними [1], близько 30 % пацієнтів із біполярним розладом та шизофренією було поставлено помилковий діагноз. Процес виявлення симптомів ментального захворювання, а також розрізнення його підтипу, є складним процесом, що потребує відповідної кваліфікації фахівця. Складовою частиною діагностування захворювань такого типу являється аналіз мовлення пацієнта. Алогія (бідність мовлення), відсутність постійного фокусу на темі мовлення, некогерентне мовлення, постійне застосування метафор можуть свідчити про наявність відповідних симптомів. Тому доцільно застосування автоматизованих методів оцінювання різних характеристик мовлення пацієнта з метою виявлення відхилень від встановлених статистичних значень. Зазначені методи відносяться до завдань напряму оброблення природної мови (Natural Language Processing, NLP). Ураховуючи відсутність уніфікованої структури тексту, наявність неоднозначності термінів залежно від контексту їх використання, завдання оброблення природної мови не можуть бути вирішені за допомогою визначеного алгоритму. Пошук закономірностей та виявлення зв'язку між елементами тексту здійснюється за допомогою методів машинного навчання: регресійні моделі, дерева рішень, глибоке навчання (багатошарові нейронні мережі). Розглянемо існуючі методи, засновані на різних методах машинного навчання, для виявлення симптомів ментального захворювання за допомогою аналізу мовлення пацієнта.

Мета роботи: здійснення порівняльного аналізу існуючих методів виявлення симптомів ментального захворювання, заснованих на використанні методів оброблення природної мови; проведення експериментальної перевірки ефективності запропонованого методу на основі аналізу зв'язності елементів тексту.

Результати та їх обговорення. Особливістю застосування методів машинного навчання, що використовуються в області оброблення природної мови, для виявлення симптомів ментального захворювання є відсутність загальнодоступної бази даних, що містить тексти пацієнтів. Формування навчальної вибірки для оптимізації параметрів моделі здійснюється дослідниками власноруч під час проведення експерименту. Таким чином зменшується розмір навчальної вибірки, що, на даний момент, унеможливлює використання моделей глибокого навчання, незважаючи на ефективність їх застосування в інших завданнях оброблення природної мови. Тому, крім досягнення необхідного значення точності у класифікації текстів на різні категорії (типи, підтипи), метою зазначених методів є виявлення властивостей тексту, що впливають на вихідне значення моделі. Відповідно до існуючих досліджень, основним критерієм вважається семантична когерентність тексту — наявність постійного фокусу навколо теми доповіді/діалогу [2]. Розглянемо детальніше існуючі методи виявлення симптомів ментального захворювання та відповідні характеристики тексту, що в них використовуються.

Модель тангенційності [3] здійснює оцінювання семантичної когерентності як міри схожості питання та відповіді пацієнта на нього, а саме, відстежує відхилення змісту висловлювань пацієнта відносно тематики відповіді. Текст відповіді розбивається на «вікна» — фрази з фіксованою кількістю слів. Далі виконується формалізоване представлення питання та «вікон» як векторів семантичного простору: міра схожості питання та «вікна» прямо пропорційна косинусній відстані між відповідними векторами. Наприклад, косинусна відстань між фразами «радіус сфери» та «діаметр сфери» дорівнює 0,6, між фразами «радіус сфери» та «сфера впливу» — 0,1. Таке представлення пропонується здійснити за допомогою латентно семантичного аналізу (Latent Semantic Analysis, LSA): окремі слова представлено у вигляді векторів семантичного простору; формалізація фраз здійснюється за допомогою усереднення значень векторів відповідних слів. В ітеративний спосіб розраховується міра схожості кожного «вікна» з питанням. На основі отриманої послідовності значень будується регресійна модель (лінійна регресія), що відображає зміну міри схожості питання та відповіді залежно від часу. Приклади такої залежності зображено на рис. 1.

На основі побудованої моделі здійснюється аналіз нахилу лінії. Автори роботи розглядають лише негативний нахил лінії: чим він крутіший, тим більше віддаляється тематика відповіді від питання. Таке віддалення може свідчити про нездатність пацієнта фокусувати свої думки («політ ідей»).

На відміну від моделі тангенційності, модель некогерентності [4] виконує аналіз лише тексту відповіді пацієнта. Основна ідея моделі некогерентності полягає у дослідженні семантичного зв'язку тексту на рівні речень. Спочатку виконується попередня обробка тексту, а саме представлення тексту як масиву речень, речень — як масиву слів; крім того, для кожного слова визначається частина

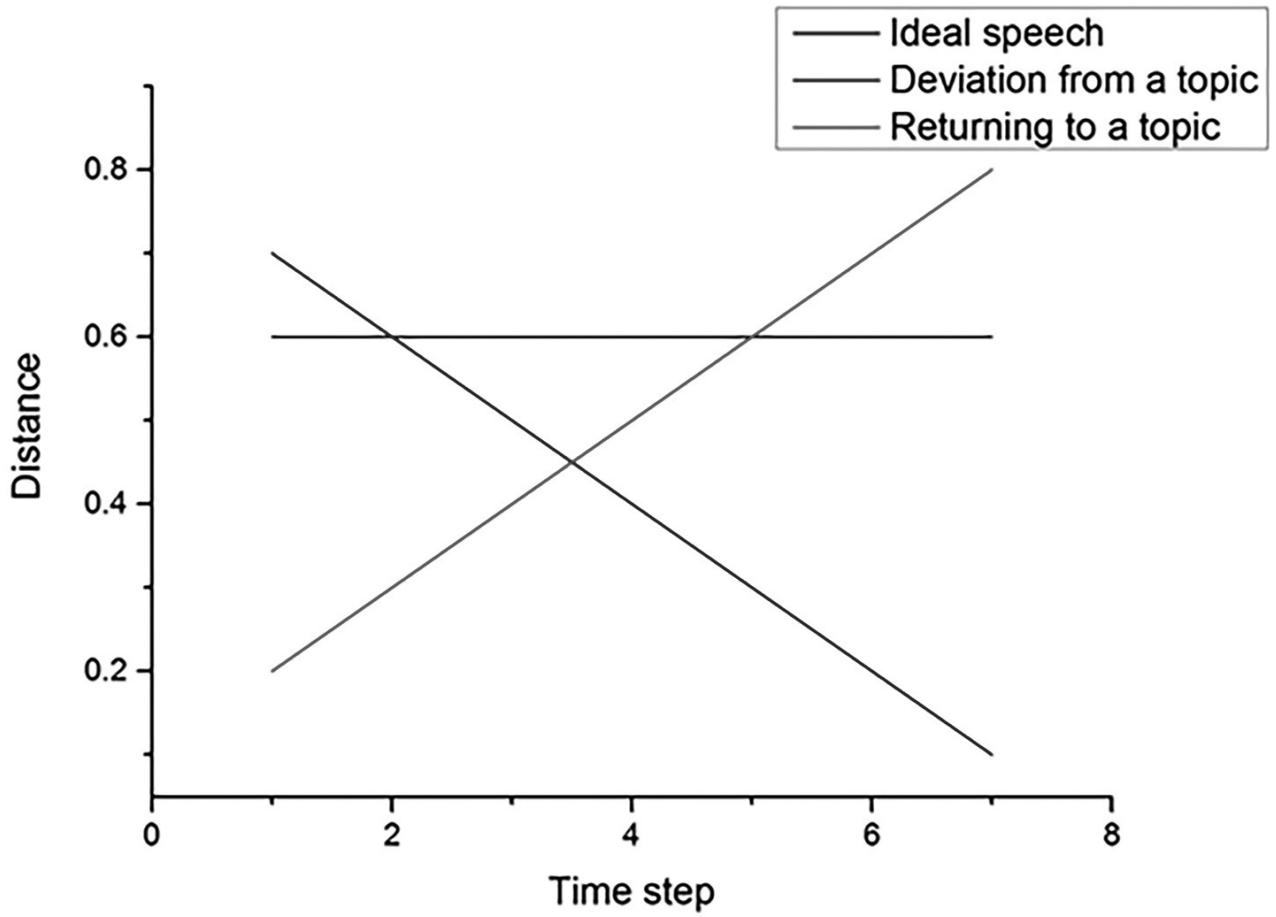

Рис. 1. Приклад побудови регресійної моделі для 3 варіантів: ідеальне мовлення (без відхилень), відхилення від теми, наближення до теми

мови для виявлення інформативних (іменник, дієслово, прикметник тощо) та допоміжних (прийменник, сполучник) елементів. Далі здійснюється векторне представлення кожного речення в спосіб, аналогічний до моделі тангенційності. Для формування вектору ознак, відповідного вхідному тексту, розраховуються такі величини:

- мінімальне значення косинусної відстані між сусідніми реченнями в межах тексту;
- максимальна довжина речень (кількість слів);
- астота вживання допоміжних елементів (відношення кількості допоміжних елементів до загальної кількості слів).

На основі сформованої вибірки здійснюється навчання бінарної моделі-класифікатора (в роботі здійснювалася побудова опуклої оболонки для розмежування текстів людей із симптомами ментального захворювання та здорових осіб).

Моделі тангенційності та некогерентності використовуються в різних дослідженнях як метрики оцінювання семантичної когерентності тексту. Варто зазначити, що точність розглянутих моделей залежить від попередньо навченої моделі семантичного представлення елементів тексту. Отже, використання інших підходів представлення тексту в семантичному векторному просторі (Word2Vec [5], Doc2Vec [6], GloVe [7], ELMo [8], BERT [9]) та навчання моделі семантичного представлення відповідно до предметної області можуть підвищити точність зазначених моделей. Однак, варто звернути увагу на наступний недолік обох моделей: постійне повторення фраз чи речень дозволить підвищити вихідну оцінку вхідного тексту. Регулярне повторення слів чи фраз може ідентифікувати симптоми захворювання, тому, навпаки, потрібно враховувати наявність постійних повторень у тексті для зменшення вихідної оцінки. В роботі [10] здійснюється порівняльний аналіз впливу різних підходів семантичного представлення тексту (LSA, GloVe, Sent2Vec, Word2Vec) на точність моделей тангенційності та некогерентності. Крім того, для врахування наявності повторень пропонується виконати нормалізацію векторів слів за допомогою додаткового застосування методів TF-IDF та SIF

(Smooth Inverse Frequency). Такий підхід дозволить зменшити вплив елементів тексту, що постійно використовуються в мовленні особи. Проте застосування TF-IDF та SIF передбачає наявність додаткового текстового корпусу: його тематика повинна відноситися до предметної області текстів, що аналізуються. Тобто, з'являється додаткова залежність точності моделей від зазначеного корпусу.

Крім моделей оцінювання семантичної когерентності, в роботі [10] пропонується розглянути вплив неоднозначних займенників на роботу класифікатору. Під неоднозначними займенниками розуміють такі елементи тексту:

- займенники, що посилаються на об'єкт, який не згадується в тексті;
- займенники, що посилаються на об'єкт, який згадується в тексті пізніше (катафора).

Найчастіше вживаними неоднозначними займенниками виявилися займенники «вони», «їх» та «він»; середня частота їхнього використання серед пацієнтів із симптомами захворювання склала 3,2 (середня довжина відповіді — 300 слів), серед здорових осіб — менше 1. Для пошуку неоднозначних займенників застосовується метод пошуку кореферентних пар [11] у тексті. На основі частоти появи неоднозначних займенників і значень моделей тангенційності та некогерентності побудовано бінарні класифікатори — random forest та логістична регресія — з метою виявлення впливу кожної з характеристик на вихідний результат. У табл. 1 наведено значимості кожної характеристики для зазначених класифікаторів [10]. Незважаючи на більшу значимість моделей некогерентності та тангенційності порівняно з відповідним показником частоти вживання неоднозначних займенників, доцільно аналізувати останню метрику для підвищення точності класифікатора.

Таблиця 1

Значимість різних характеристик моделей-класифікаторів після навчання (для random forest — важливість характеристик, для логістичної регресії — відповідні коефіцієнти)

Характеристика	Random forest	Логістична регресія
Модель некогерентності	0,44	-0,06
Модель тангенційності	0,36	-0,05
Неоднозначні займенники	0,19	0,04

Крім моделей семантичної когерентності тексту та, відповідно, різних варіантів моделей семантичного представлення тексту, в роботі [12] пропонується розглянути інші лінгвістичні характеристики: лексичне різноманіття, лексична щільність, синтаксична складність. Кожна з цих характеристик може бути представленна декількома метриками. Метрики BI (Brunet's Index) [12], MATTR [13], HS (Honore's Statistic) [14] розраховуються для представлення лексичного різноманіття тексту. Використання власне зазначених метрик було обумовлено ефективністю їх застосування для аналізу показників пацієнтів із хворобою Альцгеймера [15]. Лексична щільність [16] розраховується як відношення кількості інформативних слів (іменників, займенників тощо) до їх загальної кількості в тексті. Синтаксична складність може бути представлена як середня довжина речення чи висота синтаксичного дерева. На основі отриманих ознак було навчено класифікатори (наївний баєсів класифікатор та логістична регресія) для виконання таких розмежувань:

- здорові особи та пацієнти з симптомами шизофренії;
- пацієнти з симптоми шизофренії та особи з наявністю біполярного розладу.

Отже, на відміну від попередньо розглянутих методів, такий набір ознак тексту дозволяє виконувати не лише ідентифікацію ментального захворювання (точність 96 %), але й класифікацію хвороби (точність 82 %). Врахування зазначених лінгвістичних характеристик є доцільним для аналізу та передбачення можливих ментальних захворювань.

Розглянуті вище дослідження проводилися для англомовних осіб. У роботах [17] і [18] проводився аналіз ефективності моделей тангенційності та некогерентності для текстів німецькою та російською мовами відповідно. Модель некогерентності з застосуванням комбінації векторного представлення речень GloVe+TF-IDF виявилася найефективнішою для німецькомовного корпусу. Однак всі інші варіанти застосування моделей виявилися неспроможними виконувати класифікацію текстів.

На відміну від англомовних та німецькомовних текстів, точність моделі некогерентності для російськомовних текстів є нижчою за відповідний показник моделі тангенційності. Такі відмінності в ефективності застосування моделей для різних мов можуть свідчити про таке:

- необхідність врахування особливостей певної мови. Наприклад, у німецькій мові варто враховувати наявність написання слів із великої літери та суворий порядок слів; у російській — наявність відмінків;
- залежність точності методів від моделі семантичного представлення слів. Навчання моделей семантичного представлення елементів тексту здійснювалося на різних текстових корпусах для кожної мови, що може впливати на точність розрахунку косинусної відстані між фрагментами;
- необхідність проведення мовленнєвого діалогу з пацієнтом. На відміну від англомовних та німецькомовних пацієнтів, російськомовні записували відповідь на питання, отже, мали додатковий час для аналізу своїх думок.

Розглянувши основні моделі прогнозування симптомів ментального захворювання, засновані на методах оброблення природної мови, можна зробити висновок про доцільність додаткового аналізу мовлення пацієнта за допомогою характеристики, спільної для всіх мов. Додатково до семантичної цілісності елементів тексту пропонується аналізувати зв'язність його фрагментів за допомогою оцінювання спільних термінів і наявності кореферентних об'єктів. На відміну від катафор, що можуть свідчити про певні відхилення від теми під час процесу мислення, застосування анафор (об'єктів, що посилаються на елемент, який раніше був зазначений у тексті) дозволяє відстежувати зв'язок між усіма частинами тексту незалежно від їхнього розташування. Крім того, зазначений зв'язок кореферентних пар не залежить від особливостей певної мови, що вказує на доцільність пошуку анафор у межах тексту з подальшим аналізом зв'язку відповідних слів чи словосполучень. Аналіз зв'язності тексту пропонується виконувати за допомогою графу узгодженості словосполучень [19].

На відміну від розглянутих моделей, аналіз речень тексту $T = \{S_1, S_2, ..., S_N\}$ виконується на рівні словосполучень, а не слів чи «вікон»:

$$S_i = \{P_1, P_2, ..., P_M\}, i \in \{1, 2, ..., N\} \quad (1)$$

Такий підхід обумовлений додатковою перевіркою структурної узгодженості елементів речення та автоматизованою фільтрацією стоп-слів під час процесу екстракції словосполучень. Для оцінювання міри зв'язності двох речень S_i та S_j здійснюється побудова двочасткового орієнтованого графу K_{ij}: кожна підмножина вершин інтерпретує словосполучення відповідного речення. Крім того, вилучаються дублюючі вершини графу. У такий спосіб враховується негативний вплив постійних повторень фраз у тексті на вихідну оцінку зв'язності. Приклад такого вилучення та побудови графу узгодження словосполучень зображено на рис. 2. Значення ваг ребер графу обраховується за допомогою аналізу спільних термінів і кореферентних об'єктів; міра зв'язності двох речень оцінюється як відношення середнього значення напівстепені виходу графу узгодженості словосполучень до модуля різниці відповідних порядкових номерів речень:

$$\mathrm{Coh}(S_i, S_j) = \frac{\mathrm{avg}(\mathrm{outdegree}(K_{ij}))}{|i - j|} \quad (2)$$

Для оцінки зв'язності тексту виконується побудова повнозв'язного графу . Множина вершин відповідає множині речень; значення ребер розраховується як міра зв'язності відповідних речень. Варто зазначити, що для оцінки зв'язності тексту вирішено враховувати зв'язок між всіма його елементами, на відміну від моделей некогерентності та тангенійності. Отже, зв'язність речення розраховується у такий спосіб:

$$\mathrm{Cohesion}(T) = \mathrm{avg}(\mathrm{outdegree}(G)) \quad (3)$$

Для перевірки ефективності запропонованого підходу порівняно з моделями семантичної когерентності та іншими лінгвістичними характеристиками тексту вирішено виконати навчання класифікаційних моделей із подальшим аналізом відповідних параметрів. Англомовні інтерв'ю для підготовки навчальної вибірки отримано з робіт [20-22]. Формування векторів ознак виконано за допомогою розрахунку таких характеристик:

- зв'язність тексту, розрахована за допомогою пропонованого графу узгодженості словосполучень;
- семантична цілісність тексту, отримана за допомогою моделі некогерентності;
- метрика лексичної щільності;
- метрики лексичної різноманітності мовлення.

Після формування навчальної вибірки даних виконано навчання двох класифікаційних моделей: набору дерев рішень та логістичної регресії. У табл. 2 наведено параметри моделей після навчання: значимості характеристик для набору дерев рішень і коефіцієнти для логістичної регресії.

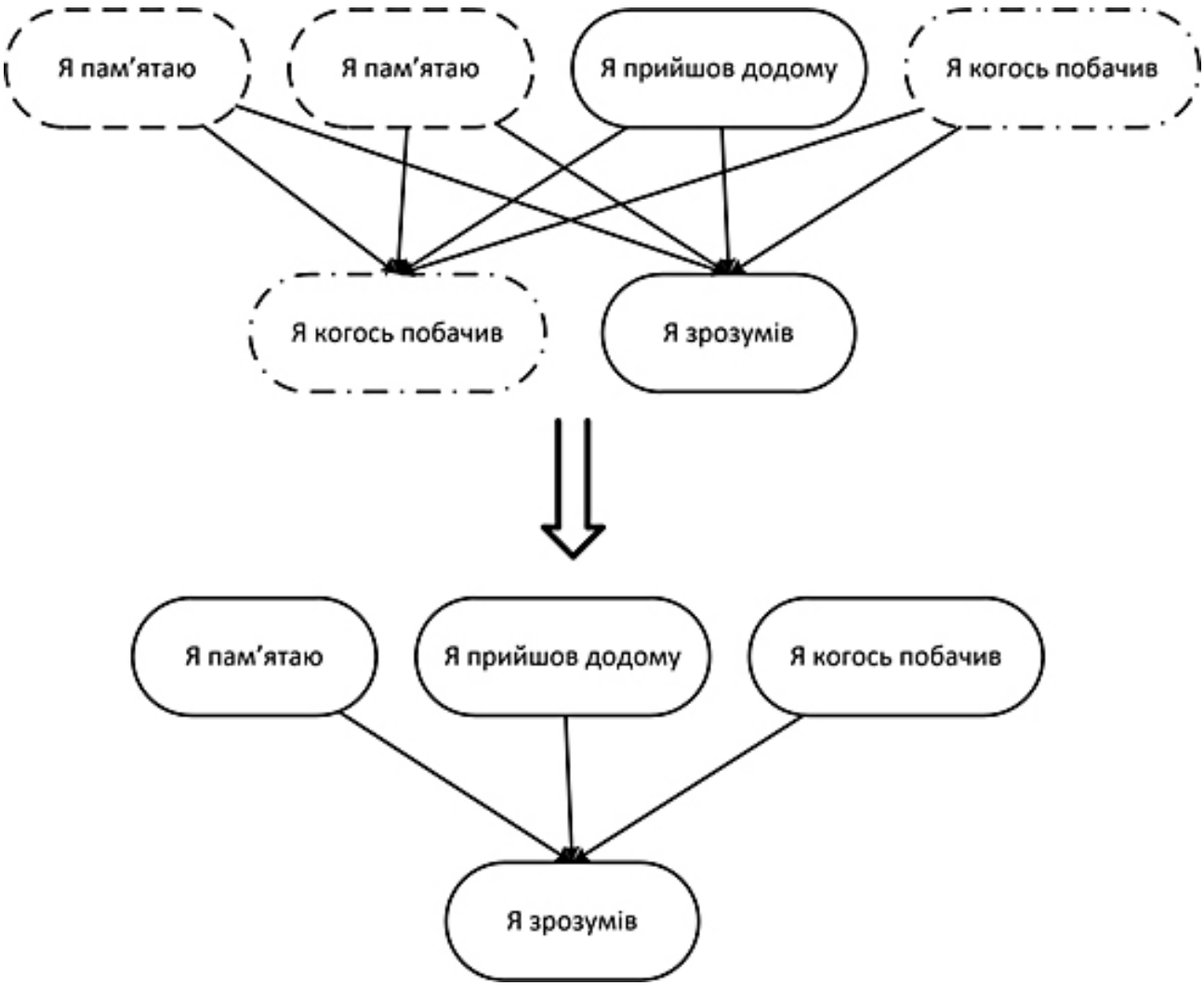

Рис. 2. Приклад побудови графу узгодженості словосполучень та вилучення словосполучень, що повторюються

Таблиця 2

Параметри класифікаційних моделей, заснованих на графі узгодженості словосполучень, після навчання: значимості характеристик для набору дерев рішень і коефіцієнти для логістичної регресії

Характеристика	Набір дерев рішень	Логістична регресія
Cohesion	0,27	0,47
FOC	0,23	1,78
FuncW	0,12	1,34
MATTR	0,18	1,32
BI	0,21	-0,30

Значення важливості метрики для набору дерев рішень є найвищим порівняно з показниками інших метрик. Отримані значення вказують на доцільність оцінювання зв'язності тексту як додаткової характеристики для побудови класифікаційної моделі. Крім того, відповідні параметри метрики можуть свідчити про необхідність аналізу семантичної когерентності тексту як основної характеристики для прогнозування симптомів ментального захворювання.

Висновки. 1. Як основний критерій відмінності текстів здорових та хворих осіб, представлені методи виявлення симптомів ментального захворювання використовують семантичну когерентність текстів. Оцінювання семантичної когерентності здійснюється на основі таких моделей: некогерентності та тангенційності.

2. Формалізацію елементів тексту в семантичному векторному просторі доцільно здійснювати з використанням комбінації різних семантичних моделей векторного представлення слів чи речень. Таких підхід дозволяє враховувати наявність повторень у тексті, характерних для людей із ментальними захворюваннями, однак, передбачає наявність додаткового текстового корпусу для розрахунку статистичних даних про елементи тексту.

3. Аналіз лінгвістичних характеристик тексту (лексична різноманітність, лексична щільність тощо) дозволяє не тільки підвищити точність класифікації текстів здорових та хворих осіб, а й здійснити розрізнення пацієнтів за типом захворювання.

4. Ефективність проаналізованих методів для текстів різних мов залежить від їхніх особливостей та точності відповідних моделей семантичного представлення елементів тексту.

5. Отримані результати експериментальної перевірки можуть свідчити про доцільність розгляду міри зв'язності тексту, заснованої на графі узгодженості словосполучень, як додаткової характеристики для побудови ефективної моделі прогнозування ментальних захворювань.

Література.

1. Altamura C. Differential diagnoses and management strategies in patients with schizophrenia and bipolar disorder / C. Altamura, J. Goikolea. // Neuropsychiatric Disease and Treatment. — 2008. — 4 (1). — P. 311-317.
2. Погорілий С. Д. Метод розрахунку когерентності українського тексту / С. Д. Погорілий, А. А. Крамов // Реєстрація, зберігання і обробка даних. — 2018. — Т. 20. — № 4. — С. 64-75.
3. Elvevåg B. Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia / B. Elvevåg, P. W. Foltz, D. R. Weinberger, T. E. Goldberg et. al. // Schizophrenia research. — 2007. — 93 (1-3). — P. 304-316.
4. Bedi G. Automated analysis of free speech predicts psychosis onset in high-risk youths / G. Bedi, F. Carrillo, G.A. Cecchi et. al. // npj Schizophrenia. — 2015. — № 1. — P. 1-20.
5. Mikolov T. Distributed representations of words and phrases and their compositionality / T. Mikolov, I. Sutskever, K. Chen et. al. // Proceedings of the 26th International Conference on Neural Information Processing Systems. — 2013. — P. 3111-3119.
6. Mikolov T. Distributed representations of sentences and documents / T. Mikolov, Q. Le // International Conference on Machine Learning. — 2014. — P. 1188-1196.
7. Pennington J. Global vectors for word representation / J. Pennington, R. Socher, C. D. Manning // Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). — 2014. — P. 1532-1543.
8. Peters M. Deep Contextualized Word Representations / M. Peters, M. Neumann, M. Iyyer et. al. // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). — 2018. — P. 2227-2237.
9. Devlin J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding / J. Devlin, M. Chang, K. Lee et. al. // Proceedings of NAACL-HLT 2019. — 2019. — P. 4171-4186.
10. Iter D. Automatic Detection of Incoherent Speech for Diagnosing Schizophrenia / D. Iter, J. Yoon, D. Jurafsky // Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic. — 2018. — P. 136-146.
11. Pogorilyy S. Coreference Resolution Method Using a Convolutional Neural Network / S. Pogorilyy, A. Kramov // Proceedings of 2019 IEEE International Conference on Advanced Trends in Information Theory. — 2019. — P. 397-401.
12. Voleti R. Objective Assessment of Social Skills Using Automated Language Analysis for Identification of Schizophrenia and Bipolar Disorder / R. Voleti, S. Woolridge, J. M. Liss et. al. // Proc. Interspeech 2019. — 2019. — P. 1433-1437.
13. Covington M.A. Cutting the Gordian Knot: The Moving-Average Type—Token Ratio (MATTR) / M. A. Covington, J. D. McFall // Journal of Quantitative Linguistics. — 2010. — 17 (2). — P. 94-100.
14. Honore A. Some Simple Measures of Richness of Vocabulary / A. Honore // Association for Literary and Linguistic Computing Bulletin. — 1979. — № 7 (2). — P. 172-177.
15. Bucks R. S. Analysis of spontaneous, conversational speech in dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance / R. S. Bucks, S. Singh, J. M. Cuerden et. al. // Aphasiology. — 2000. — № 14 (1). — P. 71-91.
16. Johansson V. Lexical Diversity and Lexical Density in Speech and Writing: A Developmental Perspective / V. Johansson // Working Papers in Linguistics. — 2009. — № 53. — P. 61-79.

17. Just S. Coherence models in schizophrenia / S. Just, E. Haegert, N. Kořánová // Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology. — 2019. — P. 126-136.
18. Panicheva P. Semantic Coherence in Schizophrenia in Russian Written Texts / P. Panicheva, T. Litvinova // 2019 25th Conference of Open Innovations Association (FRUCT). — 2019. — P. 241-249.
19. Kramov A. Evaluating text coherence based on the graph of the consistency of phrases to identify symptoms of schizophrenia / A. Kramov // Data Recording, Storage & Processing. — 2020. — № 22 (1).
20. Andreasen N. C. Scale for the assessment of thought, language, and communication (TLC) / N. C. Andreasen // Schizophrenia bulletin. — 1986. — № 12 (3). — P. 473-482.
21. Seeman M. V., Cole H. J. The effect of increasing personal contact in schizophrenia / M. V. Seeman, H. J. Cole // Comprehensive psychiatry. — 1977. — № 18 (3). — P. 283-293.
22. English Speeches with English Subtitles — English Speeches [Електронний ресурс]. — Режим доступу: https://www.englishspeecheschannel.com/english-speeches.

References.

1. Altamura, C., Goikolea, J. (2008). Differential diagnoses and management strategies in patients with schizophrenia and bipolar disorder. *Neuropsychiatric Disease and Treatment, 4 (1),* 311-7.
2. Pogoriliy, S. D., Kramov, A. A. (2018). Metod rozrakhunku koherentnosti ukrayins′koho tekstu. [Method of calculating the coherence of the Ukrainian text]. *Reyestratsiya, zberihannya i obrobka danykh (Registration, storage and data processing), 20:4,* 64-75. [In Ukrainian].
3. Elvevåg, B., Foltz, P. W., Weinberger, D. R., Goldberg T. E. et. al. (2007). *Quantifying incoherence in speech: an automated methodology and novel application to schizophrenia. Schizophrenia research, 93 (1-3),* 304-16.
4. Bedi, G., Carrillo, F., Cecchi, G. A. et. al. (2015). Automated analysis of free speech predicts psychosis onset in high-risk youths. *npj Schizophrenia, 1,* 1-20.
5. Mikolov, T., Sutskever, I., Chen, K. et. al. (2013). Distributed representations of words and phrases and their compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems, 3111*-9.
6. Mikolov, T., Le Q. (2014). Distributed representations of sentences and documents. *International Conference on Machine Learning, 1188*-96.
7. Pennington, J., Socher, R., Manning, C. D. (2014). *Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 1532*-43.
8. Peters, M., Neumann, M., Iyyer M. et. al. (2018). Deep Contextualized Word Representations. Proceedings of the 2018 Conference of the North American *Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers),* 2227-37.
9. Devlin, J., Chang, M., Lee, K. et. al. (2029). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT 2019,* 4171-86.
10. Iter, D., Yoon, J., Jurafsky, D. (2018). Automatic Detection of Incoherent Speech for Diagnosing Schizophrenia. *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic,* 136-46.
11. Pogorilyy, S. D., Kramov, A. A. (2019). Coreference Resolution Method Using a Convolutional Neural Network. *Proceedings of 2019 IEEE International Conference on Advanced Trends in Information Theory,* 397-401.
12. Voleti, R., Woolridge, S., Liss, J. M. et. al. (2019). O*bjective Assessment of Social Skills Using Automated Language Analysis for Identification of Schizophrenia and Bipolar Disorder. Proc. Interspeech 2019,* 1433-7.
13. Covington, M. A., McFall, J. D. (2010). Cutting the Gordian Knot: The Moving-Average Type—Token Ratio (MATTR). J*ournal of Quantitative Linguistics, 17 (2),* 94-100.
14. Honore, A. (1979). Some Simple Measures of Richness of Vocabulary. *Association for Literary and Linguistic Computing Bulletin, № 7 (2),* 172-7.
15. Bucks, R. S. (2000). Analysis of spontaneous, conversational speech in *dementia of Alzheimer type: Evaluation of an objective technique for analysing lexical performance. Aphasiology, 14 (1),* 71-91.
16. Johansson, V. (2018). Lexical Diversity and Lexical Density in Speech and Writing: A Developmental Perspective. *Working Papers in Linguistics, 53,* 61-79.
17. Just, S., Haegert, E., Kořánová, N. (2019). *Coherence models in schizophrenia. Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology,* 126-36.
18. Panicheva, P., Litvinova, T. (2019). *Semantic Coherence in Schizophrenia in Russian Written Texts. 25th Conference of Open Innovations Association (FRUCT),* 241-9.
19. Kramov, A. (2020). Evaluating text coherence based on the graph of the consistency of phrases to identify symptoms of schizophrenia. *Data Recording, Storage & Processing,* 22 (1).
20. Andreasen, N. C. (1986). *Scale for the assessment of thought, language, and communication (TLC). Schizophrenia bulletin, 12 (3),* 473-82.
21. Seeman, M. V., Cole, H. J. (1977). T*he effect of increasing personal contact in schizophrenia. Comprehensive psychiatry, 18 (3),* 283-93.
22. English Speeches with English Subtitles — English Speeches. URL: https://www.englishspeecheschannel.com/english-speeches.