DOI: 10.18372/2410-7840.20.12654
УДК 003.26

ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ РОБОТИ ПРОГРАМНОГО КОМПЛЕКСУ ПРОВЕДЕННЯ АТАКИ НА ЛІНГВІСТИЧНУ СТЕГОСИСТЕМУ

Ярослав Тарасенко

Стеганографія, зокрема лінгвістична набуває нових форм, зумовлених використанням комп'ютерних технологій, мережі Інтернет та зростанням впливу інформаційних технологій на всі сфери життя людини. Сучасні методи стегоаналізу текстової інформації не розраховані на ефективне використання їх для проведення атаки на лінгвістичну стегосистему, а системи автоматизованого реферування текстів не враховують можливості використання стеганографічних методів у тексті. Автором було реалізовано та попередньо протестовано програмний комплекс проведення атаки на лінгвістичну стегосистему на основі методу стиснення текстової інформації для лінгвістичної стеганографії. Проведення остаточної експериментальної перевірки працездатності та ефективності реалізації розробленого методу є актуальним. Тому в роботі уточнюються отримані попередні дані та наводяться остаточні висновки про ефективність та дієвість програмного комплексу, а відповідно і методу. Розроблено методику експерименту на основі функціонального тестування та додатковий модуль оцінки експериментальних результатів. Таким чином, отримані попередні результати та висновки були підтверджені, а виявлені недоліки виправлені. Доведено гіпотезу про можливість майже стовідоткового видалення стегоповідомлення, прихованого у тексті без втрати семантичної структури та справедливість допущення про неможливість відновлення стегоповідомлення після проведення модифікації тексту розробленим програмним комплексом. В результаті порівняння з аналогами було доведено, що хоч розроблена система стегоаналізу і демонструє гіршу ефективність, проте охоплює значно ширший спектр досліджуваних елементів, що забезпечує значно вищий показник ефективності стегоатаки стисненням. Доведено ефективність використання розробленого методу та програмного комплексу для задач стеганографії.

Ключові слова: *програмний комплекс, лінгвістична стеганографія, протидія методам стеганографії, семантичне стиснення, текстова стеганографія, автоматизований стегоаналіз, лінгвістичні системи стегоаналізу, видалення стегоповідомлення.*

ВСТУП

На сьогоднішній день, зростання впливу інформаційних технологій на всі сфери життя людини спричиняє появу нових та все більш небезпечних шляхів витоку секретної інформації. З'являються новітні методи приховування повідомлень за умови відсутності даних навіть про можливу наявність повідомлення, що може знаходитись як у звичайній діловій чи приватній переписці так і на офіційних інформаційних ресурсах. Таким чином, стародавня наука стеганографія набуває нових більш небезпечних форм, зумовлених використанням комп'ютерних технологій та мережі Інтернет. На ряду з популярним на момент написання статті використанням зображень в якості контейнеру, що може нести секретне повідомлення все більшого поширення набуває використання у цій ролі тексту природньої мови, а саме розвиваються методи комп'ютерної лінгвістичної стеганографії, що прямо впливають на пошук шляхів задоволення потреб сучасної кібербезпеки в рамках протидії цим методам, що описані в [2] та породжують появу автоматизованих засобів лінгвістичної стеганографії, стегоаналізу та проведення атак на лінгвістичну стегосистему з метою видалення, модифікації, розшифрування чи заміни стегоповідомлення у тексті. Сучасні методи стегоаналізу текстової інформації [7, 9, 10, 11] не розраховані на ефективне використання їх для проведення атаки на лінгвістичну стегосистему з метою видалення стегоповідомлення, а ефективність проведення атаки буде дорівнювати ефективності стегоаналізу. Системи ж автоматизованого реферування текстів [5, 8] не враховують можливості використання стеганографічних методів у тексті, тому проведення експериментальної перевірки працездатності та ефективності розробленого методу семантичного стиснення текстової інформації для протидії комп'ютерній лінгвістичній стеганографії та його реалізації у вигляді програмного комплексу проведення атаки на лінгвістичну стегосистему [3] з метою доведення більшої його ефективності є актуальним завданням.

Метою дослідження є проведення роботи по уточненню отриманих попередніх результатів у [3] шляхом збільшення кількості тестових наборів для отримання остаточних висновків про ефективність та дієвість методу та програмного комплексу, що реалізує атаку на лінгвістичну стегосистему, направлену на видалення стегоповідомлення шляхом семантичного стиснення з втратами та збереження початкової морфологічно-синтаксичної та семантичної структури тексту [3], а також метою є виправлення виявлених недоліків при попередньому тестуванні розробленого програмного продукту.

Крім того, необхідним є доведення гіпотези про можливість майже стовідсоткового видалення стегоповідомлення, прихованого у тексті без втрати семантичної структури та основного змісту, на ефективність якого критично не впливають такі фактори, як ентропія тексту чи його об'єм, а також доведення справедливості допущення про неможливість відновлення стегоповідомлення після проведення модифікації тексту.

МЕТОДИКА

Зазначена точність поширених алгоритмів автоматизованого стегоаналізу для: [7] – 91,175%, для [10] – становить 97,19%, для [9] – точність виявлення наявності стегоповідомлення складає 87,39%, 95,51%, 98,50%, 99,15% та 99,57% при розмірі сегменту 5кБ, 10кБ, 20кБ, 30кБ і 40кБ відповідно, для [11] точність перевищує 90%. Більш детально переваги та недоліки алгоритмів описані в [3]. В свою чергу, заявлена ефективність алгоритмі та систем стиснення тексту у [5] складає 64-71%, а у [8] – 53-70%, їх недоліки описані у [3].

В той же час, не вирішеною залишається задача експериментального дослідження розробленого комплексу, та доведення більшої ефективності використання методу порівняно з аналогами. Для вирішення цієї задачі були поставлені наступні завдання:

1. Визначення коефіцієнту семантичного стиснення текстів та ефективності проведення стегоатаки в залежності від його об'єму, ентропії, та стилю.
2. Дослідження впливу високої ентропії тексту на ефективність стиснення.
3. Визначення частоти появи помилок першого та другого роду при проведенні стегоаналізу.
4. Доведення ефективності та працездатності методу та програмного комплексу.
5. Визначення відсотку стегоповідомлення, що може бути видалене з тексту в залежності від використаного методу стеганографії.
6. Доведення справедливості допущення про неможливість відновлення стегоповідомлення після проведення модифікації тексту.

Для спрощення проведення експериментів до функціоналу програмного комплексу було додано декілька додаткових систем, реалізовних в тимчасовому модулі оцінки експериментальних результатів. По перше, мова йде про лічильник помилок першого та другого роду, що базується на можливості задання точного місцезнаходження бітів прихованого повідомлення з подальшим порівнянням із результатами проведеного стегоаналізу. По друге, систему порівняння навмисно змінених частин тексту з метою приховування стегоповідомлення з частинами тексту, що пройшли модифікацію в процесі стиснення.

Важливою також є система визначення об'єму початкового та результуючого тексту та визначення їх відсоткової разниці і збору статистики. Останньою є підсистема обліку часу виконання операцій.

Для проведення експерименту було обрано один із способів проведення тестування програмного забезпечення – функціональне тестування [4], як таке, що може одночасно задовольнити потреби тестування функціональних можливостей програмного комплексу як програмного продукту та сприятиме отриманню точних результатів дослідження застосування методу, на основі якого розроблено програмний комплекс. Згідно зі структурною моделлю проведення функціонального тестування, приведеної в [4] та адаптованої під задачі тестування конкретного програмного продукту описаного в роботі [3], модель проведення дослідження можна описати схемою (рис. 1).

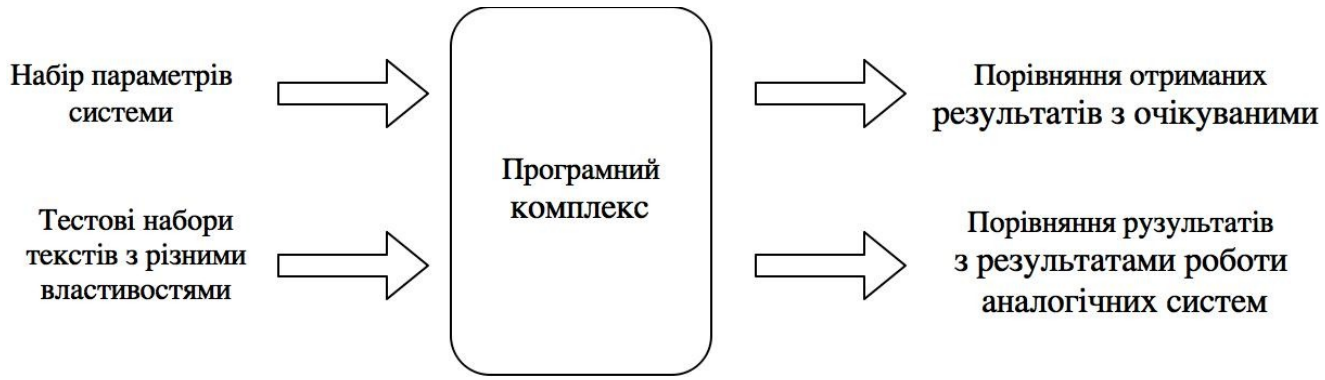

Рис. 1. Структурна схема проведення функціонального дослідження

Як видно зі схеми, програмний комплекс отримує відповідний набір параметрів, що впливають на його функціональні особливості. З кожним набором параметрів проводиться тестування,

використовуючи різні тексти з особливими властивостями, зумовлені однією з описаних причин проведення експериментального дослідження. Після обробки вхідних даних програмним комплексом отримуються певні результати, які зіставляються з очікуваними результатами роботи системи, а також з результатами роботи аналогічних систем, на основі чого робляться відповідні висновки.

Для ефективної перевірки роботоспроможності програмного комплексу було підготовлено набори тестових матеріалів для дослідження усіх функцій системи та її поведінки при виконанні кожної із зазначених завдань дослідження. Тестові набори були розподілені по групам у певній ієрархії.

Об'єм виступає головним елементом розподілу тестових наборів. Він представлений кількома групами, кожна з яких перевірялася різними наборами параметрів, що характеризують стиль тексту та його ентропію. Кожен стиль перевірявся наборами текстів, що характери-зуються наявністю чи відсутністю стегопові-домлення, а до кожного тексту, у якому було приховано стегоповідомлення застосовувались різні методи стеганографії. Як зазначається в роботі [1] в офіційно-діловому стилі адекватність ентропії принципова, а отже можна зробити висновок, що ентропія низька, оскільки можливо легко передбачити типові вирази та конструкції у тексті. Для наукового теж передбачувана, а отже також низька, для публіцистичного зазначається принциповість, проте непередбачуваність ентропії, а отже вона більша, ніж у попередніх стилів. Для художного непередбачувана і непринципова, а отже найвища серед розглянутих стилів. Проте, штучно згенерований текст буде володіти максимальною ентропією у зв'язку з відсутності семантичної цілісності. Оскільки рівень ентропії впливає на ефективність автоматизованого лінгвістичного дослідження тексту та його дискурсу, а від цього безпосередньо залежить ефективність стегоаналізу, що впливає на відсоток стегоповідомлення, який може бути видаленим (ефективність стегоатаки семантичним стисненням), ефективність роботи розробленого програмного комплексу (як стиснення тексту так і відсоток видаленого стегоповідомлення) також залежить від ентропії. Так як програмний продукт враховує усі стилі та штучно згенеровані тексти, тому необхідне дослідження стабільності та ефективності його роботи у випадку використання кожного зі стилів чи у випадку їх змішування.

Згідно з критерієм розміру, тексти були розподілені на 6 груп: до 1 Кб, від 1 до 5 Кб, від 6 до 50 Кб, від 51 до 300 Кб, від 301 Кб до 1 Мб та від 1 Мб. Таким чином визначався коефіцієнт семантичного стиснення текстів та ефективності проведення стегоатаки в залежності від його об'му.

Згідно з критерієм стилю та ентропії, тексти розподілялися на 8 груп: науковий, офіційно-діловий, публіцистичний, художній, розмовний, епістолярний, конфесійний і штучно згенерований текст. Кожна група володіла унікальною притаманною відповідному стилю ентропією. Таким чином досліджувалась ефективність стиснення та проведення стегоатаки в залежності від стилю тексту та вплив високої ентропії на ефективність стиснення.

Наявність стегоповідомлення розподіляє тексти на 2 групи, де відповідно присутнє і відсутнє приховане повідомлення. Це дозволяє визначити частоту появи помилок першого та другого роду, де приймається пустий контейнер за заповнений і навпаки та дослідити таким чином ефективність стегоаналізу.

Останній критерій розподіляє тексти згідно з методами стеганографії, що використані для приховування стегоповідомлення на 4 групи, де застосовані: методи довільного інтервалу, синтаксичні, семантичні, іноваційні медоди. Це дозволяє визначити відсоток стегоповідомлення, що може бути видалений із тексту взагалі та в залежності від використаного методу стеганографії. Крім того, виявити ефективність проведення стегоатаки в залежності від використаного методу стеганографії та об'єму тексту.

Для перевірки ефективності роботи модуля інтертекстуального дискурсного аналізу обиралися тексти з мережі Інтернет як у суцільному вигляді, так і формувалися з декількох уривків.

Генерація штучних текстів проводилася за допомогою онлайн-сервісів Random Text Generator, Malevole Text Generator, Dummy Text Generator [6]. Генератори показують задовільні результати утворення текстів, що граматично відповідають нормам, проте володіють вкрай високою ентропією. Завдяки цьому, після внесення вручну відповідних змін по приховуванню стегоповідомлення можливо перевірити ефективність протидії засобам стеганографії шляхом семантичного стиснення за умови штучної генерації тексту.

Загальна кількість унікальних текстів для перевірки перевищує 4000, загальний об'єм яких сягає 2000 Мб. Використані тексти повторювались у випадку, коли досліджувався менший об'єм із прихованим повідомленням одним і тим самим методом стеганографії.

Тестування проводились на персональному комп'ютері такої конфігурації: процесор Intel Core i5-2410M, оперативна пам'ять DDR3 об'ємом 4 Гб, жорсткий диск HDD об'ємом 750 Гб з використанням операційної системи Windows 7 Professional. Застосування оновленої конфігурації може продемонструвати кращі результати по часу виконання операцій.

ЕКСПЕРИМЕНТАЛЬНА ЧАСТИНА

В результаті проведення тестування, було виявлено коефіцієнт стиснення, що в середньому дорівнює приблизно 2,7 для окремих речень та 4,76 для суцільного тексту. Таким чином після проведення стиснення, результуюче речення в середньому дорівнювало 36% від його початкового об'єму, в той час як результуючий текст складав 21% від об'єму початкового тексту. Це зумовлено можливістю проведення скорочення тексту частинами, де видаляються абзаци та мікротеми, що володіють надлишковістю відносно головної теми. В той же час імовірність видалення стегоповідомлення в середньому складає 98,65%.

Проте рівень стиснення і ефективність проведення стегоатаки, як виявлено, залежить від стилю а відповідно і ентропії тексту. Крім того, значення відрізняються і у випадку зміни загального об'єму тексту. Значення коефіцієнту стиснення та імовірності видалення в надзвичайно малих (до 1 Кб) та надзвичайно виликих (від 1 Мб) текстах демонструють нелінійні результати, на відміну від значень при дослідження текстів від 1 Кб до 1 Мб, де при зростанні об'єму спостерігається закономірна зміна коефіцієнтів. Середні показники в текстах об'ємом 50-300 Кб зображено на рисунку 2.

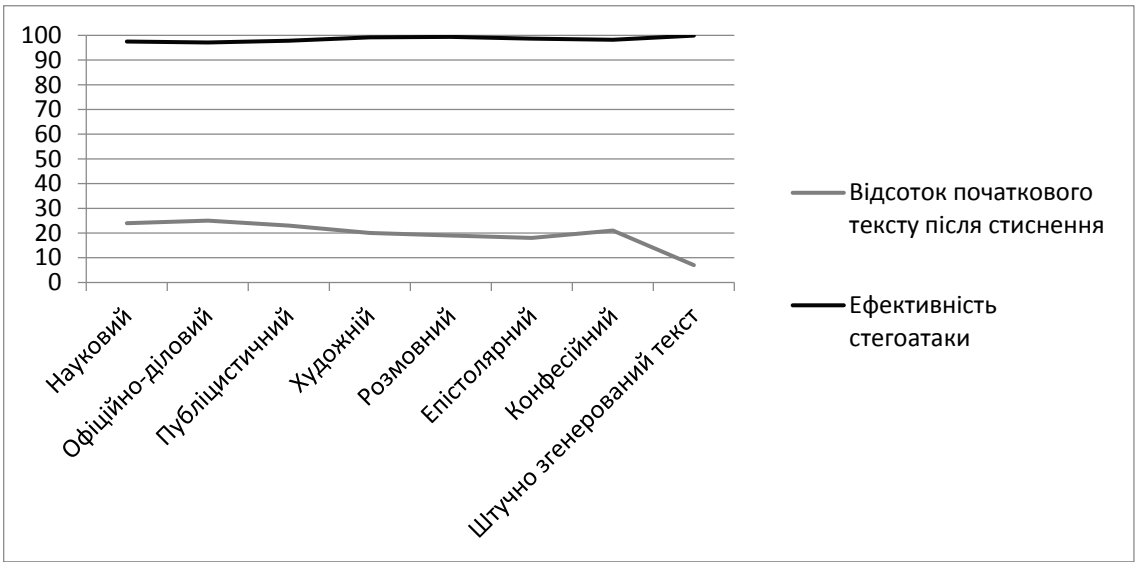

Рис. 2. Показники стиснення і ефективність стегоатаки текстів об'ємом 50-300 Кб в залежности від стилю

Як можна побачити з графіку, середній показник стиснення тексту знаходиться в межах 20% від початкого тексту та коливається в рамках 5% в залежності від стилю тексту, а відповідно і його ентропії, яка впливає на надлишковість. Якщо стилі поєднувались, як це часто буває на практиці, ефективність стиснення підвищувалась чи знижувалась відповідно до переважання елементів одного з використаних стилей.

В той же час ефективність видалення стегоповідомлення, яка визначається відсотком видаленого стегоповідомлення хоч і коливається в рамках 98,5%, проте прямо пропорційно залежить від ефективності стиснення: чим більший відсоток початкового тексту зберігає стиснений тим менша імовірність повного видалення стегоповідомлення. Цей показник в найгіршому випадку знижується до 97,1% при використанні офіційно-ділового стилю. Винятком є стиснення неосмисленого тексту, який хоч і володіє найвищою ентропією, проте показник збереженого початкового тексту не перевищує 5-7%. Середні значення ефективності стиснення та імовірності видалення стегоповідомлення по кожній групі (рис. 3) демонструє ефективність видалення стегоповідомлення у тексті будь-якого об'єму.

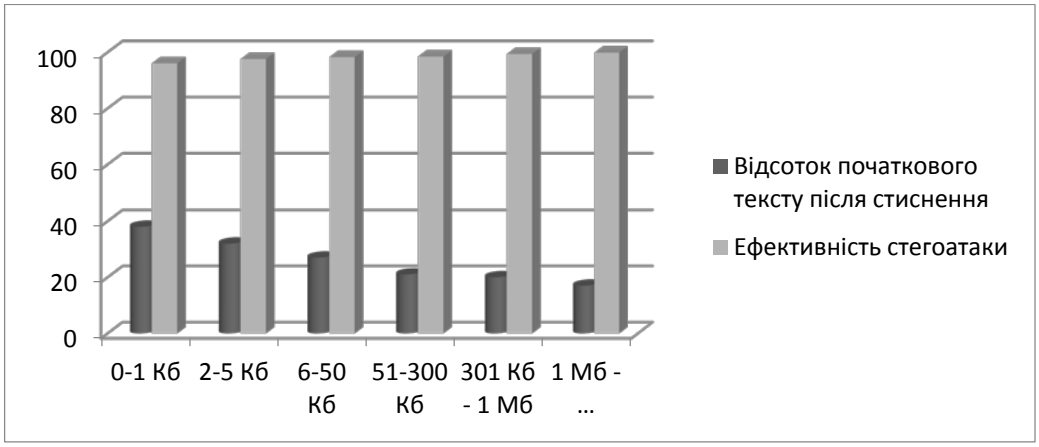

Рис. 3. Середні значення ефективності стиснення та імовірності видалення стегоповідомлення

Таким чином, можна стверджувати, що імовірність повного видалення стегоповідомлення хоч і знижується зі зниженням ефективності стиснення, яке спричинено зменшенням об'єму тексту з прихованим стегоповідомленням, проте не падає нижче показнику в 96,1%. Це доводить ефективність проведення атаки на текстову стегосистему та високу імовірність видалення стегоповідомлення, що коливається в допустимих межах при стисненні текстів різного об'єму, які володіють унікальною ентропією, а відповідно і надлишковістю.

В той же час ефективність стегоаналізу гірша, ніж у аналогічних систем, а імовірність правильного визначення наявності стегоповідомлення у тексті не перевищує 80%. Ефективність видалення стегоповідомлення забезпечується семантичним стисненням тексту у комплексі з проведеним стегоаналізом. Таким чином, якщо в аналогічних стегоаналізаторах за однакових умов ефективність складає 90% це означає, що імовірність видалення стегоповідомлення також складатиме 90% на відміну від розробленої системи, де при 80% імовірності виявлення наявності стегоповідомлення імовірність його видалення від 97% до 99,9%. В той же час, показники імовірності видалення стегоповідомлення без застосування модуля стегоаналізу складають в середньому на більше 82 %, а шанс видалення стегоповідомлення статистично має випадковий характер, що супроводжується нестабільністю проведення стегоатаки. Звідси витікає неефективність стиснення без проведення попереднього стегоаналізу.

В той час, як частота появи помилок першого і другого роду (рис. 4) збільшується зі збільшенням об'єму тексту, що приводить до зниження якості стегоаналізу, але більший об'єму тексту також зумовлює покращення стиснення, особливо завдяки системі скорочення тексту частинами, що дозволяє застосовувати засоби видалення другорядних підтем та мікротем тексту, завдяки чому зростає ефективність видалення стегоповідомлення.

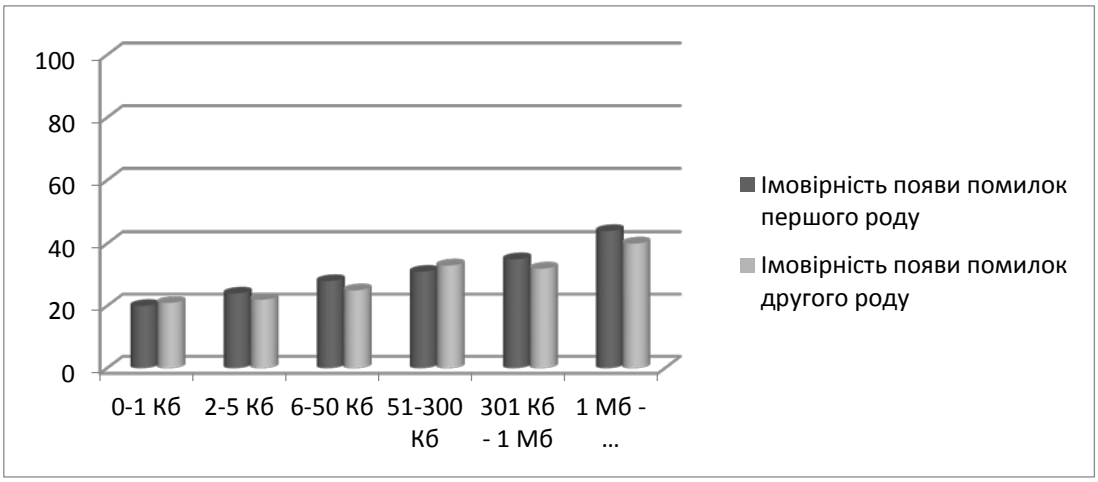

Рис. 4. Показники імовірності появи помилок першого та другого роду в залежності від об'єму тексту

Цей результат підтверджується виявленим раніше фактом, що зі збільшенням об'єму тексту збільшується коефіцієнт стиснення, а відповідно імовірність вдалого проведення стегоатаки по вида-

ленню секретного повідомлення. І, хоча ефективність стегоаналізу демонструє середні чи менші значення порівняно з аналогами, проте це компенсується ефектом стиснення, завдяки якому зростає імовірність видалення стегоповідомлення. Звідси слідує, що розроблений програмний комплекс, що практично реалізує метод семантичного стиснення текстової інформації для протидії комп'ютерній лінгвістичній стеганографії є працездатним та ефективним, проте лише завдяки використанню комплексного підходу, що реалізується завдяки гармонійній взаємодії двох невід'ємних модулів: модуля стегоаналізу та модуля стиснення на основі даних стегоаналізу.

Як видно, відсоток появи помилок першого та другого роду зменшується зі зменшенням об'єму тексту, що забезпечує ефективну взаємодію з надзвичайно малими текстами, такими як особисті повідомлення, в той же час зі збільшенням об'єму хоч і збільшується імовірність виникнення помилки при стегоаналізі, але за рахунок семантичного стиснення збільшується ефективність стегоатаки. В результаті цього забезпечується більш стабільна робота системи незалежно від об'єму тексту, що розширює можливості проведення стегоатаки та перспективи використання прогрманого комплексу.

Експериментальне дослідження також демонструє залежність ефективності роботи системи та імовірності видалення стегоповідомлення від застосованого методу стеганографії (рис. 5).

Таким чином, дослідження показало, що в середньому 98% від прихованого в тексті повідомлення з урахуванням різного об'му текстів знищується в процесі стиснення та модифікації. Звідси слідує неможливість відновлення повідомлення на стороні отримувача. У деяких випадках повідомлення видаляється на 100%. Як видно з графіку, найвищі показники ефективності можна спостерігати в найменш криптостійких методах довільного інтервалу, найнижчі у випадку використання семантичних методів у зв'язку з неможливістю видалення чи модифікації деяких важливих семантичних елементів тексту, що визначають його смислову структуру. Проте можливість приховування стегоповідомлення у частинах тексту, які не належать до надлишкових елементів надто мала, а система модифікації дозволяє змінити більшість ключових об'єктів, не впливаючи на семантичну структуру тексту. Цим зумовлюється середній показник видалення стегоповідомлення, що складає 95,6%, що забезпечує неможливість відновлення його із залишкових частин.

Рис. 5. Відсоток видаленого стегоповідомлення в залежності від застосованого методу стеганографії

На боротьбу з новітніми засоби приховування стегоповідомлення, які базуються на класичних методах стеганографії з деякими відмінностями система від початку не розрахована. Таким чином, на прикладі цих методів, було досліджено ефективність системи для роботи із засобами стеганографії, що знаходяться в розробці чи будуть розроблені в найближчому майбутньому. В результаті було виявлено, що ефективність дещо знижена та складає 97,2%. Проте показник знаходиться в межах допустимої норми. Звідси слідує, що система однаково ефективно протидіє основним наявним методам стеганографії, та зможе протидіяти методам, які з'являться в найближчому майбутньому. Крім того незначні розбіжності показнику ефективності системи для різних методів приховування повідомлення доводить ефективність системи для протидії використанню змішаних методів, що застосовується для збільшення криптостійкості стегосистеми.

Ефективність модуля дискурсного аналізу самого по собі не висока та складає 67% правильно знайдених текстів в мережі Інтернет. На це значення впливає фактор можливої зміни початкового тексту завдяки використанню синонімів чи перефразуванню тексту. Проте у випадку, коли досліджуваний текст не змінений ефективність пошуку досить висока та перевищує 90%. Хоча підсистема і не демонструє значної ефективності проте впливає на загальну картину та зумовлює

зростання показнику видалення стегоповідомлення у тексті, на що власне і напралена мета розробки методу, та застосування реалізованого на його основі програмного комплексу. Таким чином, підсистема при роботі з іншими модулями справляється із завданням покращення загального результату.

Для нейтралізації вразливості системи, описаної в [3], що полягає у збільшенні часу роботи програми зі збільшенням розміру тексту чи кількості елементів у масиві особистих не пов'язаних між собою повідомлень, було розроблено відповідні засоби протидії з метою проведення успішного тестування у стресових ситуаціях та в умовах збільшення об'єму та кількості досліджуваних текстів. Цим засобом протидії виступає підсистема аварійного відключення програми при збільшенні часу роботи до 20000 мс з подальшим записом об'єму дослідженої частини тексту чи масиву та розділення залишкового тексту чи масиву на частини, рівні цьому значенню і перезапуску системи для дослідження залишкових частин. Такий процес повторюється до моменту завершення отримання вхідних даних. Розподілення навантаження системи та швидкодії її роботи при значному зростанні трафіку, що подається програмі на дослідження та подальшу модифікацію можна спостерігати на рис. 6.

Як видно з графіку, час виконання операцій зростаючи до 20000 мс при досягенні розміру досліджуваного тексту 1 Мб падає до значень близько 3000 мс на значеннях 1,1 Мб і так до кінця дослідження тексту. Після цього дані дослідження кожного відрізку порівнюються між собою для виявлення загальної картини дослідження. Таким чином доводить ефективність роботи системи в режимі безперервної подачі вхідних даних, а відповідно і можливість використання системи на практиці при стресових навантаженнях. Під роботою системи в стресових навантаженнях розуміється випадкове чи навмисне збільшення часу виконання операцій системи у разі спроби зловмисником запобігти атаці на відправлене ним стегоповідомлення. Основний виявлений недолік програмного комплексу при попередньому тестуванні виправлено. А отже, система є стійкою при використанні у подібних умовах.

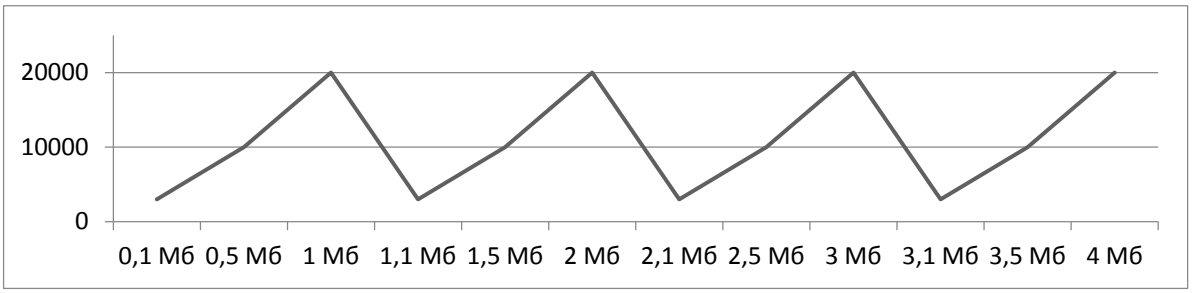

Рис. 6. Графік навантаження системи при безперервній передачі вхідних текстових даних

Таким чином, програмний комплекс демонструє ефективну роботу не залежно от обраного стилю, а також взаємодіє як з критично малими текстами (особисті повідомлення) так і з великими масивами даних. Розроблене програмне забезпечення демонструє стабільну та ефективну роботу за різних умов та в стресових ситуаціях.

РЕЗУЛЬТАТИ ТА ОБГОВОРЕННЯ

Ефективність роботи методу можна повноцінно оцінити лише в порівнянні ефективності ключових систем розробленого програмного комплексу та існуючих аналогічних систем. Перш за все необхідно порівняти результати стегоаналізу з аналогічними стегоаналізаторами (табл. 1), що використовують різні підходи дослідженя тексту.

Точність стегоаналізу та ефективність проведення стегоатаки розраховані як середнє арифметичне від значень аналізу та атаки текстів різного об'єму. Стабільність відображає яким чином змінюється ефективність аналізу зі збільшенням об'єму тексту. У випадку висхідної зміни ефективності, імовірність виявлення стегоповідомлення зростає зі збільшенням об'єму тексту, при низхідній навпаки.

Таблиця 1

Порівняння розробленої системи стегоаналізу тексту з аналогічними системами

Система	Ефективність стегоатаки	Точність стегоаналізу	Використаний підхід	Стабільність
Розроблена автором [3]	98,65%	74%	Дискурсний аналіз	Універсальна
[7]	91,175%	91,175%	Статистичний	Висхідна
[9]	96%	96%	Статистичний	Висхідна
[10]	97,19%	97,19%	Статистичний	Висхідна
[11]	90%	90%	Дослідження ентропії	Висхідна

Як можна побачити, аналогічні системи текстового стегоаналізу, що основуються на одному з представлених підходів, демонструють ефективність до 97,19%. Усі розглянуті аналогічні системи володіють висхідною ефективністю. Це означає, що ефективність стегоаналізу окремих речень чи текстів вкрай малого розміру переважно на низькому рівні, оскільки не можливо дослідити статистичні чи частотні зміни. В той же час, розроблений програмний комплекс направлений на вирішення задачі стагоаналізу та проведення атаки на стегосистему, в якій використовуються тексти вкрай малого розміру. Крім того, більшість оглянутих систем стегоаналізу володіють вузькою направленістю. Натомість запропонований стегоаналізатор має широкий спектр аналізу та дозволяє виявляти сліди використання основних відомих методів, що забезпечує ефективне проведення атаки на стегосистему, яка основується як на використанні будь-якого відомого методу стеганографії чи їх комбінації, так і методів що будуть з'являтися в найближчому майбутньому. Логічно, що зі зростанням спектру аналізу знижується ефективність виявлення стегоповідомлення, проте стегоаналізатор необхідний лише як основа для проведення стегоатаки семантичним стисненням, а заявлених 74% достатньо, щоб направити процес стиснення тексту на необхідні частини. Якщо проводити стегоатаку, базуючись на аналогічних системах стегоаналізу, то ефективність стегоатаки буде дорівнювати ефективності стегоаналізу. Крім того, ефективність буде знижуватись з використанням декількох систем у комплексі для розширення спектру стегоаналізу, або витрачати більше часу на проведення аналізу окремо кожною системою, ризикуючи отримати більше помилок першого чи другого роду.

Якщо ж говорити про ефективність стиснення розробленою системою в порівнянні з аналогічними системами реферування (табл. 2), що у зв'язку з відсутністю повноцінних програмних продуктів по реалізації атаки на стегосистему шляхом семантичного стиснення методом дискурсного аналізу можуть бути застосовані для видалення стегоповідомлення, то вони хоч і володіють певними перевагами, проте непрофільність спричиняє низьку ефективність при застосуванні їх для подібних атак.

Таблиця 2

Порівняння розробленої системи (модулю) стиснення тексту із системами реферування

Система	Рівень стиснення	Використаний метод	Врахування стеганографії
Розроблена автором [3]	64-79%	Дискурсний аналіз	Присутнє
[5]	64-71%	Дискурсний аналіз	Відсутнє
[8]	53-70%	Abstractive summarization	Відсутнє

Оскільки усі розглянуті системи не мають функції стегоаналізу, їх ефективність для задоволення потреб стеганографії досить низька, оскільки, припинення використання модулю стегоаналізу спричинює можливість відновлення стегоповідомлення після стиснення. Також показник імовірності повного видалення стегоповідомлення, на відміну від розробленої системи є характеристикою не стабільною, випадковою та залежною від використаних методів приховування стегоповідомлення, що зумовлює неможливість прогнозування ефективного видалення стегоповідомлення.

Крім того, розроблена система дає змогу підвищити ефективність стиснення за умови низької імовірності втрати семантичної структури та важливих семантичних елементів досліджуваного тексту. На основі характеристики використаного підходу можна зробити висновок, що аналогічні системи рідко володіють надійними засобами захисту початкової семантики тексту. Хоча, система [5] в повній мірі враховує семантику, оскільки застосовується дискурсний аналіз, може використовуватись для проведення стегоатаки, однак у зв'язку з тим, що в ній, як і в усіх інших, відсутнє врахування засобів стеганографії, результат стегоатаки передбачити важко. Проте, як видно з таблиці 2, ефективність стиснення за допомогою розробленої системи та запропонованої в [5] в середньому близькі, однак за рахунок застосування інтенсіональної логіки та відмінного підходу до проведення дискурсного аналізу вдалось середні значення стиснення збільшити на 8%.

Розглянуті системи реферування зустрічають труднощі при скороченні як надзвичайно малих (стиснення коротких речень є малоефективним у зв'язку з неможливістю отримати частотно-статистичну характеристику) так і надзвичайно великих текстів (порушується семантика, втрачається точність передачі початкової інформації). Запропонована розробка ефективно проводить стиснення як коротких речень так і великих текстів, що розширює спектр можливого використання методу та програмного комплексу, реалізованого на основі

цього методу для проведення стегоатаки на текстову стегосистему шляхом семантичного стиснення, основаному на дискурсному аналізі. Крім можливості стиснення, система надає можливість модифікації стисненого тексту, що відсутня в більшості систем реферування. А система врахування початкової семантики забезпечує ефективний захист від імовірності втрати важливих структурних частин тексту, що в поєднанні з системою модифікації дозволяє досягти майже 100% імовірності видалення стегоповідомлення з тексту.

ВИСНОВКИ

Таким чином, в результаті збільшення кількості та різноманітності тестових наборів вдалось отримати фінальні результати, що дали змогу зробити остаточні висновки щодо ефективності методу та, реалізованого на його основі програмного комплексу. Розроблений програмний продукт було порівняно з існуючими аналогами Для тестування було реалізовано 3 додаткові тимчасові підсистеми, об'єднані модулем оцінки експериментальних результатів.

Таким чином, остаточне тестування довело ефективність та працездатність методу і програмного комплексу для подальшого його використання на практиці. Було виявлено, що ефективність стиснення хоч і залежить від об'єму і стилю тексту, а відповідно і його ентропії, але відхілення не значні, а рівень стиснення знаходиться в межах від 64% для речень до 79% для осмислених текстів та 93% для неосмислених текстів. В той же час імовірність видалення стегоповідомлення в середньому складає 98,65%.

Імовірність виникнення помилок першого та другого роду знаходиться в межах від 20% до 44% в залежності від об'єму тексту, однак це не впливає на ефективність проведення стегоатаки, оскільки відсоток видаленого стегоповідомлення є не нижче за 95,6% та наближається до 100% в залежності від використаного методу стеганографії, що доводить припущення про неможливість відновлення стегоповідомлення після проведення процесу стиснення та модифікації тексту, а також доводиться гіпотеза про можливість майже стовідсоткового видалення стегоповідомлення, прихованого у тексті без втрати семантичної структури та основного змісту.

Крім того, було усунено недолік програмного комплексу, пов'язаний з вразливістю системи до випадкового чи навмисного перевантаження і збільшення часу роботи при аналізі великих масивів даних завдяки розробленій системі захисту від навмисного чи випадкового перевантаження програми, що забезпечує обробку безперервного потоку вхідних даних.

В результаті порівняння розробленої системи з аналогічними системами текстового стегоаналізу та реферування було доведено, що хоч розроблена система стегоаналізу і демонструє гіршу ефективність, проте охоплює значно ширший спектр досліджуваних елементів, що забезпечує значно вищий показник ефективності стегоатаки, аніж при використанні аналогічних стегоаналізаторів. Порівняння з існуючими системами реферування довело ефективність використання розробленої системи для задач стеганографії, а також виявило більший показник стиснення тексту за рахунок використання іншого підходу в дискурсному аналізу, що базується на застосуванні елементів інтенсіональної логіки та комплексного підходу до аналізу.

На основі цього можна стверджувати, що попередні результати були підтверджені. Відсоток видаленого стегоповідомлення знаходився в межах 97 ±2 %, а коефіцієнти стиснення в межах 2,5 ± 0,3 для окремих речень та 4,6 ± 0,2 для тексту.

ЛІТЕРАТУРА

[1] Н. Валгина, *Теория текста*, М.: Логос, 2003, 280 с.

[2] І. Федотова-Півень, Я. Тарасенко "Шляхи задоволення потреб сучасної кібербезпеки в рамках протидії методам комп'ютерної лінгвістичної стеганографії", *Безпека інформації*, №23(3), С. 190-196, 2017.

[3] Я. Тарасенко, "Прграмний комплекс проведення атаки на лінгвістичну стегосистему", *Безпека інформації*, №24 (1), С. 56-61, 2018.

[4] Different Types of Software Testing And How It Improves The Software Quality [Електронний ресурс]. Режим доступу: https: // www. spaceotechnologies. com/different-types-of-software-testing.

[5] E. Günes, B. Radev, "LexRank: graph-based lexical centrality as salience in text summarization", *Journal of Artificial Intelligence Research,* vol. 22, i. 1, pp. 457-479, 2004.

[6] K. Gaines, "15 Dummy Text Generators You Should Know", [Електронний ресурс]. Режим доступу: https://www.webdesignerdepot.com/2012/03/15-dummy-text-generators-you-should-know.

[7] P. Meng, L. Hang, Z. Chen, Y. Hu, W. Yang, "STBS: A Statistical Algorithm for Steganalysis of Translation-Based Steganography", *12th International Conference «Information Hiding»*, Calgary, Canada, June 28-30, Vol. 6387, pp. 208-220, 2010.

[8] R. Paulus, C. Xiong, R. Socher, "A Deep Reinforced Model for Abstractive Summarization", *arXiv preprint arXiv:1705.04304*, May 2017 [Електронний ресурс]. Режим доступу: https://arxiv.org/abs/1705.04304.

[9] Z. Chen, L. Huang, Z. Yu, L. Li, W. Yang, "A Statistical Algorithm for Linguistic Steganography Detection

Based on Distribution of Words", *Third International Conference on Availability, Reliability and Security*, Barcelona, Spain, March 04-07, pp. 558-563, 2008.

[10] Z. Chen, L. Huang, Z. Yu, W. Yang, L. Li, X. Zheng, X. Zhao, "Linguistic Steganography Detection Using Statistical Characteristics of Correlations between Words", *10th International Workshop «Information Hiding»*, Santa Barbara, USA, May 19-21, Vol. 5284, pp. 224-235, 2008.

[11] Z. Chen, L. Huang, Z. Yu, X. Zhao, X. Zhao, "Effective Linguistic Steganography Detection", *8th International Conference on Computer and Information Technology Workshops*, Sidney, Australia, July 08-11, pp. 224-229, 2008.

ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ РАБОТЫ ПРОГРАММНОГО КОМПЛЕКСА ПРОВЕДЕНИЯ АТАКИ НА ЛИНГВИСТИЧЕСКУЮ СТЕГОСИСТЕМУ

Стеганография, в частности лингвистическая приобретает новые формы, обусловленные использованием компьютерных технологий, сети Интернет и ростом влияния информационных технологий на все сферы жизни человека. Современные методы стегоанализа текстовой информации не рассчитаны на эффективное использование их для проведения атаки на лингвистическую стегосистему, а системы автоматизированного реферирования текстов не учитывают возможности использования стеганографических методов в тексте. Автором было реализовано и предварительно протестировано программный комплекс проведения атаки на лингвистическую стегосистему на основе метода сжатия текстовой информации для лингвистической стеганографии. Проведение окончательной экспериментальной проверки работоспособности реализации разработанного метода является актуальным. Поэтому в работе уточняются полученные предварительные данные и приводятся окончательные выводы об эффективности и действенности программного комплекса, а соответственно и метода. Разработана методика эксперимента на основе функционального тестирования и дополнительный модуль оценки экспериментальных результатов. Таким образом, полученные предварительные результаты и выводы были подтверждены, а выявленные недостатки исправлены. Доказана гипотеза о возможности почти стопроцентного удаления стегосообщения, скрытого в тексте без потери семантической структуры и справедливость допущения о невозможности восстановления стегосообщения после проведения модификации текста разработанным программным комплексом. В результате сравнения с аналогами было доказано, что хоть разработанная система стегоанализа и демонстрирует худшую эффективность, однако охватывает значительно более широкий спектр исследуемых элементов, обеспечивает значительно более высокий показатель эффективности стегоатаки сжатием. Доказана эффективность использования разработанного метода и программного комплекса для задач стеганографии.

Ключевые слова: программный комплекс, лингвистическая стеганография, противодействие методам стеганографии, семантическое сжатие, текстовая стеганография, автоматизированный стегоанализ, лингвистические системы стегоанализа, удаление стегосообщения

EXPERIMENTAL RESEARCH OF THE SOFTWARE COMPLEX TO ATTACK THE LINGUISTIC STEGOSYSTEM

Steganography, in particular linguistic steganography, acquires new forms due to the use of computer technologies and the Internet as well as the growing influence of information technologies on all spheres of human life. Modern methods of the textual information steganalysis are not intended for the purpose of efficiently use to attack the linguistic stegosystem, and the systems of automated text summarization do not take into account the possibility of steganographic techniques usage. The author implemented and pre-tested the software complex to attack the linguistic stegosystem on the basis of the method of the textual data compression for linguistic steganography. So, the final experimental research of the developed method implementation`s effectiveness is relevant. Therefore, the preliminary obtained data are specified in the work and the final conclusions about the effectiveness of the software complex and, accordingly, the method are presented. The experiment's method based on the functional testing has been chosen and an additional module for evaluating experimental results has been developed. Thus, the preliminary results and conclusions were confirmed and the identified defects were corrected. The hypothesis about the stegomessage hidden in the text removal possibility by almost one hundred percent without loss of the semantic structure and the validity of assumptions about the impossibility of recovering the stegomessage after the modification of the text by the developed software complex is proved. As a result of comparison with analogues, it was proved that even though the developed system of steganalysis has shown a worse efficiency, it covers much wider range of investigated elements, which provides a significantly higher efficiency of the attack based on compression. The efficiency of using the developed method and software complex for steganography problems is proved.

Keywords: software complex, linguistic steganography, counteracting the methods of steganography, semantic compression, textual steganography, automated steganalysis, linguistic systems of steganalysis, removal of hidden message.

Тарасенко Ярослав Володимирович, аспірант кафедри інформаційної безпеки та комп'ютерної інженерії, Черкаський державний технологічний університет.
E-mail: yaroslav.tarasenko93@gmail.com.

Тарасенко Ярослав Владимирович, аспирант кафедры информационной безопасности и компьютерной инженерии, Черкасский государственный технологический университет.

Tarasenko Yaroslav, postgraduate student of the department of information security and computer engineering, Cherkassy state technological university.