УНИВЕРЗИТЕТ „ГОЦЕ ДЕЛЧЕВ“ – ШТИП

ФАКУЛТЕТ ЗА ИНФОРМАТИКА

Катедра за компјутерско инженерство и
комуникациски технологии

Иван Стојанов

КРИЕЊЕ НА ПОДАТОЦИ ВО ЕЛЕКТРОНСКИ ДОКУМЕНТИ

– МАГИСТЕРСКИ ТРУД –

Штип, декември, 2020 година

Комисија за оценка и одбрана

Ментор:	Доц. д-р. Александра Милева *Факултет за информатика* Универзитет „Гоце Делчев“ – Штип
Член:	Доц. д-р. Наташа Стојковиќ *Факултет за информатика* Универзитет „Гоце Делчев“ – Штип
Член:	Доц. д-р. Доне Стојанов *Факултет за информатика* Универзитет „Гоце Делчев“ – Штип

Членови на Комисијата за оценка и одбрана

Претседател:	Доц. д-р. Доне Стојанов *Факултет за информатика* Универзитет „Гоце Делчев“ – Штип
Член:	Доц. д-р. Наташа Стојковиќ *Факултет за информатика* Универзитет „Гоце Делчев“ – Штип
Член:	Доц. д-р. Александра Милева *Факултет за информатика* Универзитет „Гоце Делчев“ – Штип

Научно поле:	Информатика
Научна област:	Дигитална стеганографија и стеганализа
Датум на одбрана:	01.12.2020

Листа на рецензирани и објавени трудови произлезени од истражувањето

1. *I., Stojanov, A. Mileva, I. Stojanovic: "A New Property Coding in Text Steganography of Microsoft Word Documents," in Securware 2014: The Eighth International Conference on Emerging Security Information, Systems and Technologies, pp.25 - 30, 2014.*

2. *I., Stojanov, A. Mileva, D. Stojanov, N. Stojkovik: "MSWordSST - A New Steganalytical Tool for Microsoft Word Documents," in 12th ICT Innovations Conference, 2020 (web-proceeding).*

Криење на податоци во електронски документи

Краток извадок

Во овој труд читателот ќе има можност да се запознае со концептот за криење на податоци во електронски документи (претежно во *MS Word* документи), како и со концептот за детектирање на потенцијално постоење на скриени пораки во нив. По извршеното истражување и презентирање на постоечките техники за процесите за криење на податоци, дополнително се разгледуваат техники кои се предложени и имплементирани од страна на авторот. Во вториот дел од трудот се опфатени постоечки техники за детекција на потенцијално криење на податоци и разгледување на соодветните имплементации за истите, од страна на авторот.

Клучни зборови

стефанографија, стеганализа, MSWord, криење на информации

Data hiding in electronic documents

Abstract

In this paper, the author presents the concept of embedding secret messages into electronic documents (mostly refer to *MS Word* documents) and presents the concept of detection of potential existence of hidden messages in a given document. After analyzing the results from the research and presenting a list of existing techniques used for the process of embedding the secret messages, the author gives an overview to a list of new techniques, developed and implemented for the purpose of this paper. In the second part of the paper, the author presents a list of existing techniques for detection of secret messages into electronic documents and gives an overview of the corresponding implementation for them.

Keywords

steganography, steganalysis, MSWord, information hiding

Содржина

Листа на слики

Листа на табели

1 Вовед

Комуникацијата претставува процес за разменување на информации и како таква, таа е дел од секојдневното опкружување на човекот. Тоа е социјална вештина која што претставува неизбежен сегмент од човековото битие и е дел од процесот за извршување дури и на најосновните егзистенцијални човекови потреби. Се јавува во различни облици на вербална комуникација (орална, пишана) и невербална комуникација (изрази на лице, говор на тело итн.), но без разлика на нејзиниот облик, секогаш крајната цел е лицето кое што ја иницира да испрати информации до друго лице / или до одредена дестинација.

Приватната комуникација помеѓу луѓето се одвива преку приватни канали, до кои само засегнатите лица имаат пристап. Идејата за испраќање на приватни информации од едно лице на друго, при користење на јавен канал, ја наметнува потребата од изнаоѓање на начини за криење на тие приватни информации во оригиналната испратена порака.

Времето на дигитализацијата во кое живееме, ни овозможува пристап до најразлични алатки, платформи и начини за комуникација, кои до пред две декади воопшто и не постоеле. Во овој труд ќе се задржиме на комуникацијата која што е во дигитален формат и која што ги искористува компјутерите и новите информациски технологии поврзани со нив. Секој електронски документ, како на пример: слика, текстуален документ, аудио, видео, дел од мрежен протокол итн., може да биде модификуван на начин што во себе ќе содржи скриена порака која на прв поглед не е видлива за крајниот корисник. Лицето кое го испраќа документот, ја вметнува скриената порака, документот е достапен за поголем број на корисници, од кои само оние коишто знаат за постоењето на скриената порака можат да ја детектираат и прочитаат.

Самиот концепт за криење на податоци е чувствителен и сам по себе претставува закана за сајбер-безбедноста, со оглед на тоа дека може да биде користен за лоши цели од страна на терористи или криминалци. Со оглед на тоа дека за секој метод кој е користен за криење на пораки (т.н. кодер), постои соодветен спротивен метод за читање на скриените пораки (т.н. декодер) и со самото тоа дека постои значителен број на вакви методи и техники, речиси е невозможно да се изнајде начин по секоја цена да се прочита скриената порака, од страна на лице кое не е запознаено со користениот метод (во овој случај лице

кое се грижи за безбедноста). Токму поради тоа се јавува потребата за потенцијална детекција на документите кои во себе содржат скриени пораки. Детекцијата се врши преку анализа на својствата на документот, за да се добие заклучок за тоа дали содржината на документот е сомнителна или не. Самата детекција не ја чита скриената порака, туку како резултат ја дава веројатноста за нејзиното постоење. Оваа анализа, исто така се темели на определени методи и веројатноста која ја пресметува претставува веројатност токму за методите што се испитуваат. Негативниот резултат за постоење на скриена порака за еден метод, не значи дека во документот не постои скриена порака со друг метод. Токму поради обемноста на методите за криење и обемноста на методите за детекција, конкретното поле од науката за вгнездувањето и детекцијата на скриените пораки е сеопфатно и комплексно.

Од другата страна пак, позитивната страна на концептот за криење на пораки е тоа што може да има примена при следење на документи, примена на дигитални печати, заштита на авторски права, автентикација и слично [2] [3] [4]. Во оваа примена, испраќачот го искористува концептот и преку скриена порака верификува дека тој е всушност авторот на документот, преку вметнување на свој потпис и подоцна споредба на истиот.

1.1 Преглед на литературата

Овој труд е конципиран на начин што откако е даден краток вовед во првата глава и откако е разгледана користената литература во втората глава, се пристапува кон одредување на целите на истражувањето во третата глава.

Во четвртата глава подетално се разгледани постоечките методи за стеганографија и постоечките методи за стеганализа.

Четвртата глава е поделена на начин што најпрво се разгледуваат стеганографските методи базирани на формати кои вршат поместување на линиите и зборовите [8] [12] [13], ги менуваат карактеристиките на знаците [12] [13] и нивната модулација на осветленоста [18], вршат манипулации со празните места преку отворените методи [16] [17] или пак манипулации со невидливи знаци [21] [22] [23] [24] [25], ги искористуваат типовите на фонтови на знаците [28] или пак нивните кодови [29].

Разгледани се и лингвистичките стеганографски методи за следење на промените на зборовите [31], за мапирање на зборовите [32] и за менување на нивното спелување [34].

Од стеганографските методи базирани на генерирање, разгледан е пример за генератор на шеги [35] и пример за генератор на листи [39].

Потоа се пристапува кон разгледување на методите за стеганализа кои се базирани на врските помеѓу зборовите [41] или преку дефинирање на одредени т.н. „мета-својства“ на содржината и нивна анализа [42]. Стеганализата може да се врши и преку конволуција на семантичкиот простор [44], да се базира на форматите на фонтовите [45] или да се анализира степенот на поместување на зборовите [46].

Во петтата глава детално се разгледуваат нови предложени методи за стеганографија [48] и за стеганализа [56] од страна на авторот, додека пак во шестата глава се разгледува имплементацијата на предложените методи, преку две посебни алатки.

1.2 Цел на истражувањето

Целта на истражувањето е детално да се анализира постоечката литература за различни типови на техники и различни начини за криење на податоци во електронски документи (претежно *Miscrosoft Word* документи). Согласно резултатите од анализираните податоци, да се пристапи кон предлагање и креирање на нови техники кои освен теоретски, ќе бидат имплементирани и практично во некој програмски јазик. Ова ќе значи изнаоѓање на нови начини за криење на податоци (скриени пораки) во веќе постоечки документ од страна на испраќачот, но соодветно на тоа и начини за детекција и читање на скриените пораки од страна на примачот. По креирањето на новите техники и нивната практична имплементација, да се изврши испитување и споредување на новите техники со веќе постоечките, во однос на неколку параметри, како капацитетот на вгнездување и дополнителната меморија која тие ја додаваат во однос на оригиналниот документ. Овој дел од трудот би резултирал со алатка во која што крајниот корисник ќе може да избере документ и да вгнезди скриена порака кој самиот ја внесува (преку делот за кодирање). Алатката би имала соодветен интерфејс за кодирање (вгнездување) и декодирање (читање). Делот за декодирање би поддржувал внесување на документ и соодветно читање на

скриената порака од него, под претпоставка дека корисникот е запознаен со тоа, кој метод бил користен во процесот на кодирањето.

Од аспект на безбедноста, како дел од целиот процес за криење на податоци во електронски документи, истражувањето ќе го опфати и делот за анализа на документи со цел детекција за тоа дали документот е потенцијален носител на скриена порака. И овој дел е фокусиран на *Miscrosoft Word* документи и притоа се врши анализа на постоечките техники за ваков вид на детекција. Соодветно на новите техники за криење на податоци имплементирани во првиот дел, се развиваат и практично се имплементираат нови техники за детекција на нив. По нивната имплементација, се врши испитување за тоа колку време е потребно за целиот процес на детекција да даде одредени резултати, во однос на големината на документот и во однос на користените техники. Крајната цел во овој дел би била изработка на алатка која што би детектирала кои техники потенцијално би можеле да се искористени врз документот. Со анализа на својствата на документот и преку нивна споредба со својствата на техниките за криење на податоците, се добиваат соодветни резултати, преку кои корисникот самиот може да заклучи колкава е веројатноста врз документот да е применета некоја од техниките. Оваа алатка ќе биде проширена со имплементација за детекција и на други дополнителни техники (кои ќе бидат анализирани при истражувањето), се со цел крајниот корисник да може да детектира поголем број на можни сценарија.

2 Методи на истражувачката работа

2.1 Постоечки методи за стеганографија

Стеганографија (*steganography*) претставува процес на вгнездување на скриени пораки во документи, при што документите навидум изгледаат сосема нормално и крајните корисниците не се ни свесни за постоењето на скриената пораката во нив. Целта на вгнездувањето на скриените пораките е тие да се префрлат до одредена цел (да допрат до одредени корисници), а притоа да се пренесат по најнормален пат, па до нив ќе има пристап голема маса на корисници. Бидејќи документите може да допрат до поголем број на корисници т.е. пренесување на скриена порака преку јавен канал, техниките за стеганографија се стремат кон тоа незабележително да ја менуваат оригиналната содржина додека во неа ја вгнездуваат скриената порака.

Самиот термин стеганографија е многу поширок концепт за праќање на скриени пораки и тој се јавувал во најразлични форми низ историјата. Името доаѓа од латинскиот збор „*steganographia*“, кој претставува комбинација од грчките зборови „*steganós*“ - што значи прикриен и *„-graphia*“ - што значи пишување. Терминот за прв пат се спомнува во 1499 година од страна на Johannes Trithemius во неговата книга „*Steganographia*“[1], а првата забележана примена датира уште од 440 години пред нашата ера кога Херодот пишувал за тоа како грчкиот владетел Хистијај му пратил порака на својот вазал Аристагора, преку најверниот слуга, со тоа што му ја избричел главата, напишал порака на кожата на главата на слугата и кога косата на слугата пораснала, тој го пратил кај вазалот со инструкции за тоа како да ја види скриената порака. Опишаниот пример претставува типичен пример за тоа како стеганографијата функционира во праксата: додека слугата го минува патот од испраќачот до примачот, тој доаѓа до допир со голема маса на луѓе, коишто не се ни свесни за скриената пораката што тој ја пренесува.

Стеганографијата како идеја за испраќање на скриени пораки, наоѓа примена во најразлични сфери од општеството и истата зазема различни облици. Низ историјата постојат и други забележани примери за пренос на скриени пораки, кои можат да се класифицираат како облици на стеганографија:

- Скриени пораки на хартија напишани со невидливо мастило, кои можат да бидат прочитани само со користење на УВ зраци од коишто мастилото свети;

- Скриени пораки напишани на пликоа, во делот кој е прикриен со печатот;
- Методот за шифрирање „*null chipper*“ [5] е еден вид на стеганографија со тоа што скриената порака е дистрибуирана низ оригиналната содржина која што навидум изгледа сосема нормално и не е сомнителна. Скриената порака е распределена во помали делови, според одредено правило (зборови, букви и сл.) и таа лесно се декодира доколку се знае применетото правило за кодирање, на пример: со спојувањето на секоја N-та буква од секој збор, спојувањето на првите букви од секој N-ти збор, спојувањето на секој N-ти збор во дадена реченица и сл.;
- Пораки напишани со користење на Морзеов код врз платно, кое што потоа е ткаено како дел од облека која ја носи соодветен курир;
- Во 1966 година, на телевизиско гостување на кое што бил присилен, како американски затвореник од војната, Џеремаја Дентон [6] непречено трепкал со очите и преку Морзеов код ја праќал пораката „тортура“ и со тоа им дал до знаење на Американците, дека американските затвореници се измачувани во затворите во Северен Виетнам.

Модерната дигитална стеганографија, која што ја знаеме денес, се појавува околу 1985 година, со појавата на персоналните компјутери кога почнува и примената на класичните стеганографски проблеми. Иако развојот на стеганографските техники и примени на почетокот бил бавен, со текот на времето се развиваат разни облици кои ги искористуваат новите технологии:

- Вгнездување на податоци во најмалите битови на слики и аудио датотеки кои веќе претходно содржат шумови, на начин што новите битови поминуваат незабележително како дел од веќе постоечките шумови;
- Невлијателно менување на звукот на аудио документи или менување на аудио сигнали;
- Вгнездување на слики како рамки во видеоматеријал, кои опционално се прикажуваат при пуштање на видеото во режим кој користи побрза или побавна брзина;
- Вгнездување на порака во рамките на шифрирани податоци/случајно генерирани пораки, со тоа што пораката која се крие е самата шифрирана и се запишува врз самиот блок на шифрирани/случајно генерирани пораки. Крајниот корисник кој не е свесен за постоењето на скриените пораки т.е. податоците кои

тој моментално ги гледа за него претставуваат шифра/или случајно генериран блок кои очекувано се во формат кој не е читлив, без самиот да знае за постоењето на скриена порака во самата таа нечитлива шифра;

- Вметнување на невлијателно задоцнување при испраќање на пакети преку мрежата, при клик на тастатурата. Ваквиот вид на задоцнување на пакетите преку мрежата, може да се искористи за на пренос скриени/шифрирани податоци;
- Менување на редоследот на елементите во множество од пакети низ мрежата;
- Додавање на невлијателни податоци во главата (header) на пакетите, при нивниот пренос преку мрежен протокол;
- Вметнување на податоци во секции од документот кои се игнорирани од страна на системот, на пример при постоење на условно извршување на кодот, во делот во кој сме сигурни дека системот нема да се изврши.

Секој начин за вгнездување на скриена порака во оригинална содржина, се класифицира како метод/техника за стеганографија. Самото сознание дека оригиналната порака во себе содржи скриена порака, не е доволно за декодирање на неа. Само оние засегнати страни кои ја знаат користената техника со која што иницијално е скриена пораката, можат да ја декодираат истата со користење на соодветна техника за декодирање која што ги извршува спротивните акции од техниката за кодирање.

Иако терминот стеганографија често се споредува со терминот криптографија, овие две техники имаат една главна разлика. Предноста на стеганографијата е тоа што засегнатите страни кои го примаат документот, не се свесни за постоењето на скриената порака и сама по себе не привлекува никакво внимание. Од друга страна пак, применувањето на криптографијата и шифрирањето на пораките, без разлика на тоа колку многу се безбедни и непробивни применетите методи, самите тие пораки привлекуваат внимание и како такви можат да се инкриминирачки во каналите за комуникација каде што шифрирањето е нелегално. Додека криптографијата претставува практика за криење на самата содржина на оригиналната порака, стеганографијата е концентрирана кон тоа до го прикрие фактот дека во оригиналната порака е вгнездена скриена порака.

Во зависност од типот на ентитетот кој се користи за пренос на скриените пораки, постојат повеќе видови на стеганографија како: стеганографија во дигиталните медиуми, стеганографија во системот со датотеки (filesystem), мрежна стеганографија, текстуална стеганографија итн.

Текстуалната стеганографија е еден од најкористените облици на стеганографија и всушност претставува криење на пораки во текст и во текстуални документи. Текстот е еден од најстарите медиуми за криење на податоци, па пред дигитализацијата и пред појавата на дигиталната стеганографија, како ентитети за пренос биле користени букви, книги, телеграми и сл. По појавата на дигиталните медиуми, текстот останува најмногу распространет дигитален медиум кој се појавува во форма на списанија, книги, веб-страници, изворен код (*source code*), договори, реклами итн. и како таков, ја наметнува текстуалната стеганографија како најпопуларен облик во оваа област.

Microsoft Word е најпопуларен софтвер за креирање, уредување и процесирање на текстуални документи и е дел од пакетот *Microsoft Office*. Неговата популарност кај просечниот корисник се должи на едноставноста за користењето на текстуалните документи како и достапноста до голем спектар на алатки за менување и уредување на стилот на документот (фонт, боја, маргини, слики, табели и сл.). Токму големиот спектар на алатки за форматирање на документот, придонесува за можноста за развивање на голем број на техники за криење на пораки во овој тип на документи, кои директно ги искористуваат овие алатки за незабележително менување на својствата на документот.

Популарноста на текстуалната стеганографија, во комбинација со популарноста на софтверот *Microsoft Word* како најкористен софтвер за текстуалните документи, го фокусира овој труд кон криењето на податоци во текст во софтверот за процесирање на текст *Microsoft Word*.

Постојат неколку главни категории на текстуална стеганографија [7]: методи базирани на формати или структурни методи [8], лингвистички методи [9] [10] и методи базирани на случајно и статистичко генерирање [11].

2.2 Стеганографски методи базирани на формати

Методите за стеганографија базирани на формати, генерално ги форматираат и ги менуваат својствата на постоечкиот текст за да се вгнездат

скриените пораки во него, без притоа да ги менуваат самите зборови или реченици. Тие ги искористуваат различните алатките за форматирање на текстот (кои се достапни во *Microsoft Word*) со цел вметнување на определена шема за кодирање, преку менување на својствата (фонтот / големината / бојата / позадината / маргините итн.) на ентитетите (буквите / броевите / зборовите / линиите / реченициите / параграфите итн.). На страната на примачот, се применува соодветна шема за детектирање на промените на својствата и секоја промена на својствата која што одговара на шемата, е потенцијален преносител на битови со податоци.

Со оглед на тоа дека има можност за промена на различни својства на различни ентити во рамките на еден документ, резултира со можноста за комбинација на повеќе методи базирани на формат во рамките на истиот тој документ. Тоа придонесува кон уште поголем број на постоечки методи базирани на формати и придонесува кон комплексноста со која тие можат да бидат креирани. Во зависност од користените техники, овој вид на методи може да поминат незабележително од страна на крајниот корисник, но да бидат лесно забележани од страна на компјутерот и обратно.

Токму методите базирани на формати се во фокусот при креирање на новите техники за стеганографија подолу, со оглед на тоа дека истражувањето во трудот се темели врз анализа на постоечките техники и искористување на достапните *Microsoft Word* алатки за форматирање на текст, со цел предлагање на нови методи за форматирање.

Главен недостаток кај методите базирани на формати е тоа што можат да бидат компромитирани од страна на просечниот корисник кој и покрај тоа што не мора да е свесен за нивното постоење, може лесно да врши промени врз документот кои директно се одразуваат врз користените шеми за кодирање и декодирање. Со едноставна промена на форматот на документот, форматот кој е користен за скриената порака може повеќе да не е валиден. На пример, доколку методот базиран на формат, го менува типот на фонтот на одредени знаци во документот (со тоа што секој тип на фонт значи пренос на одредени битови), доколку корисникот сака да примени нов фонт врз документот, со едноставна селекција на текстот и промена на неговиот фонт, логиката од страна на испраќачот е поништена. Токму поради тоа, овие методи се користат

при комуникација кога документот директно се праќа од испраќачот до примачот, без притоа да патува низ споредни канали и биде подложен под ризик да некој друг изврши промена на форматите, пред тој да биде доставен до крајната цел.

Секој метод, врши пренос на битови искористувајќи одредени својства на одредени ентитети, како знаци, реченици, параграфи итн. Во зависност од својствата и од ентитетите, секој метод има различен капацитет за пренос на скриени пораки. Со оглед на тоа дека секој знак е со големина од еден бајт (осум бита) и со оглед на тоа дека скриените пораки се пренесуваат во бинарен јазик, за пренос на порака од N знаци е потребно да се вгнездат N*8 бита. Капацитет на еден метод се дефинира преку максималниот број на битови кои можат да се сокријат користејќи го тој метод.

Во продолжение подетално се разгледани некои типови на постоечки методи базирани на формати.

2.2..1 Метод за поместување на линии

Методот за поместување на линии [8] [12] [13] (*Line Shifting*) ги користува линиите во секој *Microsoft Word* документ, како ентитети кои го вршат преносот на скриените пораки. Како својство на ентитетите/линиите се разгледува нивната позиционираност, во однос на нивната стандардна позиција. Линиите во текстот се поместуваат вертикално (нагоре или надолу) за мало незначително растојание кое не е лесно забележливо за човековото око.

Ова растојание достигнува најмногу до 1/1300 инчи нагоре или надолу од позицијата каде што линијата би била позиционирана доколку документот не би содржел скриена порака. Поместување на линијата нагоре може да преставува бинарна единица, поместувањето на линијата надолу може да преставува бинарна нула и обратно. Дополнително, се јавува потреба од постоење на т.н. контролни линии кои го дефинираат стандардното растојание помеѓу линиите и во зависност од нив, се согледува дали линиите кои се поместени всушност се поместени нагоре или надолу. Во постоечките техники, како контролни линии претежно се користат непарните линии, а поместувањето се врши на парните линии. Контролните линии се потребни при процесот на декодирање т.е. читањето на скриената порака, со тоа што при читањето на документот, се споредува растојанието помеѓу секоја парна линија со нејзините непарни

соседни линии, за да се детектира кое растојание е помало т.е. на која страна е поместена самата таа линија.

Предност на овој метод е тоа што тој наоѓа примена - како во електронски формат, така и во печатен формат. Процесот на читање на пораката во печатениот е идентичен како и електронскиот, со тоа што крајниот корисник е тој што треба да ја анализира печатената верзија со цел да ги детектира поместувањата на линиите.

Со оглед на тоа дека секое поместување на позицијата означува бинарна вредност од единица или нула, може да се констатира дека капацитетот на методот за поместување на линиите, зависи од бројот на линии кои ги има документот *Microsoft Word*, на начин што две линии се преносители на само еден бит. За криење на еден знак, користејќи го методот за поместување на линиите, е потребно документот да содржи најмалку 17 линии т.е. 8 линии за пренос на битовите и 9 контролни линии за детекција на поместувањето.

2.2..2 Метод за поместување на зборови

Методот за поместување на зборови [12] [13] (*Word Shifting*) како ентитети за пренос на скриените пораки ги искористува зборовите во *Microsoft Word* документите, а како својство на зборовите кое е искористено при преносот е нивната хоризонтална позиционираност. Зборовите се поместуваат во лево или во десно, за незначително растојание од околу 1/150 инчи и секое поместување претставува една бинарна вредност.

Поместувањето се врши за секој парен збор, а секој непарен збор се користи како контролен збор за мерење на растојанието на поместените зборови. Текстот може да се разгледува како множество од блокови од по три збора, каде што првиот и третиот збор се контролните зборови и служат за детекција на поместувањето на вториот збор. Бидејќи растојанието помеѓу зборовите во оригиналниот документ не е стандардизирано (може да варира од збор до збор), за детекција и декодирање на скриената порака самиот кодиран документ не е доволен т.е. потребен е и оригиналниот документ заедно со кодираниот, со цел точна детекција на поместувањата предизвикани при процесот на вгнездување на пораката.

Капацитетот на методот за поместување на зборовите, е дефиниран преку бројот на зборови кои ги содржи документот и секои два збора се преносители на еден бит. За криење на еден знак (8 бита) потребни се најмалку 17 збора во документот т.е. 8 преносители и 9 контролни.

Со оглед дека двата методи за поместување се слични и користат различни ентитети како преносители на битови, тие можат да се комбинираат [8] на начин што секоја парна линија претставува ентитет за пренос (преку вертикално поместување на позицијата), а истовремено е и поделена на три блокови од зборови, каде што средниот блок е поместен на лево или на десно. Друга комбинација од двата методи [14] е кога линијата е поделена во сегменти од последователни зборови и секои два соседни сегменти споделуваат ист збор. Со поместување на зборовите во средниот сегмент, може да се кријат еден или два бита по сегмент, во зависност од статистичката дистрибуција на растојанието помеѓу зборовите во оригиналниот документ.

2.2..3 Метод за менување на карактеристиките на знаците

Методот за менување на карактеристиките на знаците [12] [13] (*Feature Coding / Character Coding*) ги менува карактеристиките само на некои одредени знаци, како што се висината, нивната позиција релативно на позицијата на другите знаци и слично. Неколку конкретни примери кои се применуваат како дел од овој метод се: издолжување или скратување на хоризонталната линија во знаците „t“ и „f“, издолжување или скратување на горниот дел од вертикалната линија во знаците „h“, „b“ и „d“, зголемување или намалување на големината на точката во знаците „i“ и „j“ и слично. Последната техника со зголемување или намалување на точката во знаците може да има голема примена во случај на користење на арапската/персиската азбука [15], со тоа што може да се примени дури на 14 знаци од истата.

Секоја промена на карактеристиките (издолжување / скратување / зголемување / намалување) која што е различна од стандардната, може да дефинира пренос на еден бит, чија што вредност се дефинира според тоа дали знакот е издолжен или скратен. Капацитетот на овој метод се дефинира преку бројот на знаци во текстот, кои се одбрани дека ќе се користат како преносители, т.е. на пример, доколку се користи само техниката со знакот „t“ тогаш

капацитетот ќе зависи само од тоа колку пати овој знак се појавува во текстот. Со оглед на тоа дека секој знак, преносител, врши пренос на еден бит, за криење на еден знак од скриената порака, се потребни 8 знаци-преносители во рамките на оригиналниот текст.

2.2..4 Метод за модулација на осветленоста

Идејата зад методот за модулација на осветленоста (*Luminance Modulation Coding*) е да се квантизира бојата т.е. интензитетот на осветленоста, на таков начин што човековото сетило за вид нема да ја согледа разликата помеѓу оригиналната и квантизираните содржина.

Авторите во [18] ја вгнездуваат скриената порака со промена на интензитетот на осветленоста индивидуално на секој ентитет за пренос, на начин што на иницијалната темна боја додаваат некоја минимална вредност (од претходно дефинирано множество од вредности со кардиналност S), при што со промената на осветленоста на секој ентитет врши пренос на $log_2 S$ битови. Како ентитет за пренос се предлага секој знак, секоја линија, одреден симбол и сл. Капацитетот на методот директно зависи од тоа кој тип на ентитет ќе биде избран како преносител на скриените битови.

Методот предложен во [19] има слични карактеристики како претходно опишаниот метод, со тоа што овде е дефинирано дека како ентитети за пренос се користат знаците и како нијанси на интензитетот на осветленоста се користат само две вредности од кои едната е светла а другата е темна. Светлата нијанса се користи за пренос на бинарна нула, а темната нијанса се користи за пренос на бинарна единица. Интензитетот на осветленоста се менува на начин што човековото око на прв поглед не може да ги препознае извршените промени и истот се менува само на букви т.е. не се менува интензитетот на осветленоста на инструкциските знаци, со цел да има подобра детекција при процесот на декодирање.

Во [20] е претставена поинаква верзија за модулација на осветленоста, каде што текстуалниот документ се скенира и вгнездувањето на скриената порака се врши врз копија од скенираниот документ. Врз сликата се врши сегментација, поделба на блокови и идентификација на пиксели, со тоа што промената на интензитетот на осветленоста се врши само врз одредени пиксели. Како таков,

и покрај тоа што станува збор за текстуален документ, овој формат на стеганографија не е погоден за манипулација и детекција на скриени пораки во софтверот *Microsoft Word*.

2.2..5 Отворени методи

Во групата на отворени методи (*Open Spaces / White Stag*) спаѓаат повеќе техники кои вршат вметнување на невидливи специјални знаци (претежно празни места) во оригиналниот документ. Ваквото вметнување на празни места може да се врши на крајот на секоја реченица, на крајот на секоја линија, помеѓу зборовите [16], на крајот на секој параграф [17] итн. Основните типови на отворени методи се базираат на техниката за вметнување на едно или две празни места, во зависност од тоа дали со вметнувањето се кодира бинарна нула или бинарна единица соодветно и постојат т.н. три основни типови на отворени методи.

Првиот основен тип на отворен метод е кога вметнувањето на празни места (едно или две, во зависност од тоа дали се кодира бинарна нула или бинарна единица) се применува на крајот од секој знак кој означува премин во нова линија. Со оглед на тоа дека секој знак за премин во нова линија вообичаено претставува завршување на еден параграф и започнување на друг, капацитетот на овој тип на метод зависи од бројот на параграфи кои ги има во документот.

При вториот тип на отворен метод, претходно се дефинира максималниот број на празни места кои можат да се внесат на крајот на секоја линија, на начин што две празни места се користат за кодирање на еден бит по линија, четири празни места се користат за кодирање на два бита по линија, осум празни места се користат за копирање на три бита по линија, итн. Од дефинираниот максимален број на празни места, зависи капацитетот на методот т.е. бројот на битови кои можат да се сокријат, но според тој број, зависи и мапирањето според кое бројот на празни места се мапира во битови од скриената порака. Во продолжение се дадени табелите за мапирање, во случај на дефиниран максимален број на две, четири и осум празни места. Во случај кога мапираната вредност е нулта вредност, тогаш соодветната линија не е преносител на битови од скриената порака.

Табела 1: Мапирање на битови со отворен метод - линии, при дефиниран максимален број на две празни места

број на празни места	мапирана вредност
/	*null*
1	0
2	1

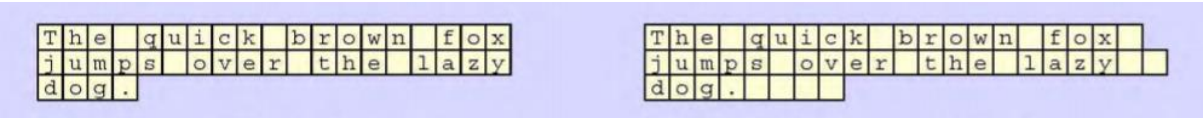
Табела 2: Мапирање на битови со отворен метод - линии, при дефиниран максимален број на четири празни места

број на празни места	мапирана вредност
/	*null*
1	00
2	01
3	10
4	11

Табела 3: Мапирање на битови со отворен метод - линии, при дефиниран максимален број на осум празни места

број на празни места	мапирана вредност
/	*null*
1	000
2	001
3	010
4	011
5	100
6	101
7	110
8	111

На *Слика 1* е даден пример за додавање на празни места на крајот на секоја линија. Доколку максималниот број на празни места е дефиниран на четири (според *Табела 2*), скриената порака во овој пример би била „*00 01 11*“, додека пак доколку максималниот број на празни места е дефиниран на осум (според *Табела 3*), скриената порака во овој пример би била „*000 001 011*“.



Слика 1: Основен тип на отворен метод - линии

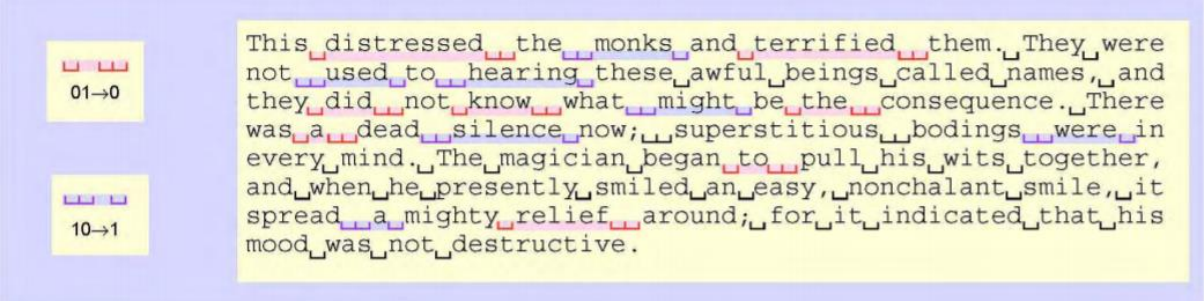
Капацитетот на овој метод директно зависи од бројот на линии во документот, со тоа што секоја линија претставува ентитет за пренос, а индиректно зависи од дефинираниот максимален број на празни места т.е. се влијае на тоа колку битови се сокриени во секоја линија која што претставува ентитет за пренос.

Во третиот тип на отворен метод вметнувањето на празните места се врши после секој збор и при истиот тој метод се јавува потреба од воведување на дополнителна логика при кодирањето, со оглед на тоа дека по крајот на битовите наменети за скриената порака (под претпоставка дека едно празно место помеѓу зборовите претставува бинарна нула и две празни места претставуваат бинарна единица) секое празно место ќе биде лажно детектирано како бинарна нула и покрај тоа што скриената порака е веќе завршена. За таа цел се воведува дополнително кодирање, на начин што битовите детектирани како резултат од користење на едно или две празни места, дополнително се групираат во групи од по два последователни битови и преку овие групи се врши мапирање дадено во *Табела 4*.

Табела 4: Мапирање на битови со отворен метод - зборови

битови детектирани со празни места	мапирана вредност
00	null
01	0
10	1
11	null

Начинот на кој функционира вгнездувањето на скриените битови е даден во *Слика 2*, каде што јасно се гледа како при читањето на пораката, исчитаните битови се групираат во групи по два детектирани бита и потоа како тие соодветно се заменуваат со нивната мапирана вредност.



Слика 2: Основен тип на отворен метод - зборови

Доколку мапирана вредност е нулта вредност, соодветната група која што резултира со нулта вредност, не претставува преносител на скриен бит. Со самото ова, секој две последователни единични празни места, не се детектирани како преносители на бит и со самото тоа се елиминира можноста за лажна детекција на бинарни нули, по завршувањето на скриената порака. Овој метод овозможува распределба на скриените битови насекаде низ документот т.е. не е задолжително скриената порака да мора да почнува од самиот почеток на документот и да мора да се искористува секој нареден ентитет – преносител, се до моментот на нејзиното завршување (во овој случај секое празно место помеѓу два соседни зборови се јавува како ентитет – преносител).

Во примерот од *Слика 2*, детектирани групи со црвена боја резултираат со бинарни нули, а детектирани групи со виолетова боја резултираат со бинарни единици. Комбинацијата на сите добиени мапирани вредности, ја формираат скриената порака која што е вгнездена во документот. Капацитетот на третиот тип на отворен метод е поголем од претходните два, со оглед на тоа дека истиот зависи од бројот на зборови во еден документ т.е. за криење на еден бит се потребни три збора (две последователни празни места помеѓу зборовите). Во овој случај пораката не се добива директно од бројот на празните места помеѓу зборовите, туку се добива од последователната промена на бројот на празните места помеѓу соседните зборови.

2.2..6 Методи со манипулација на невидливи знаци

Методот [21] како преносител на скриените битови ги искористува невидливите знаци во документот (празно место, знак за премин во нова линија, *tab* знак и сл.) како ентитети врз кој се врши промена на бојата, па со оглед на тоа дека тие се невидливи, фактот дека имаат зададено боја не е видлив за

крајниот корисник. Секоја боја е дефинирана преку нејзините *RGB* вредности и токму овие вредности се директниот преносител на скриените битови. *RGB* вредност на боја [22] претставува комбинација од три параметри (*red, green, blue*) каде што секој параметар е вредноста на интензитетот на некоја од боите црвена, зелена или сина соодветно на скала од 0 до 255 т.е. во бинарна вредност, на скала од 00000000 до 111111111. Со ова, секој параметар е преносител на осум бита, една боја содржи три параметри, значи промената на бојата на едно празно место претставува пренос на 24 бита.

На пример, за опишување на начинот за пренос на скриената порака: 1010110101110101100011010100110011101001001110010100111001010101100010 10 истата ќе биде претставена преку повеќе групи од по осум бита, кои во продолжение соодветно се прикажани со *bold* и *italic* стилови на фонтот: **10101101***01110101***10001101***01001100***11101001***00111001***01001110***01010101***10001010**. Овие групи, претставени со децимални броеви се дадени во *Табела 5*:

Табела 5: Бинарни во децимални вредности, за дадениот пример за методот за манипулација на невидливи карактери

бинарна вредност	децимална вредност
10101101	173
01110101	117
10001101	141
01001100	76
11101001	233
00111001	57
01001110	78
01010101	85
10001010	138

Добиените децимални вредности, дополнително групирани во кластери (супер-групи) од по три елементи може да се претстават како

{173, 117, 141}, {76, 233, 57}, {78, 85, 138}

каде што секој кластер формира соодветна *RGB* вредност, на начин што секоја вредност на параметрите во еден кластер преставува *R*,*G* или *B* вредност

{*R* = 173, *G* = 117, *B* = 141}, {*R* = 76, *G* = 233, *B* = 57}, {*R* = 78, *G* = 85, *B* = 138}

и добиените вредности се всушност боите кои треба да се искористат за обојување на првите три невидливи знаци.

Во продолжение се дадени деталите за алгоритамот кој се користи при процесот на криење на битовите со манипулација на невидливите знаци, преку соодветни чекори, под претпоставка дека веќе се знае низата од бинарни единици и нули – скриената порака што треба да се сокрие. Алгоритамот е всушност детален опис на примерот коишто веќе беше разгледан претходно а истиот е даден подолу на *Слика 3*.

***Чекор 1*:** Почнувајќи од десно кон лево, низата од единици и нули се дели во групи од по осум бита. Доколку последната група од битови (онаа на најлевата страна) е група која што има помалку од осум бита, на почетокот се додаваат бинарни нули, со цел групата да се доведе до големина од осум бита. Дополнително, доколку бројот на вака добиените групи е број кој што не е делив со три, на почетокот (од лева страна) се додаваат цели групи составени од бинарни нули, се додека не се добие број на групи кој е делив со бројот три.

Чекор 2: Секои три групи од по осум бита, дополнително се групираат во кластери (супер-групи) од по три групи.

***Чекор 3*:** Секоја бинарна вредност во групите од по осум бита, се заменува со нејзината соодветна децимална вредност.

***Чекор 4*:** За првиот невидлив знак во документот се користи првиот кластер (супер-група).

***Чекор 5*:** Како вредност на *R*-параметрот (од *RGB* бојата) се поставува првата децимална вредност од соодветниот кластер (супер-група).

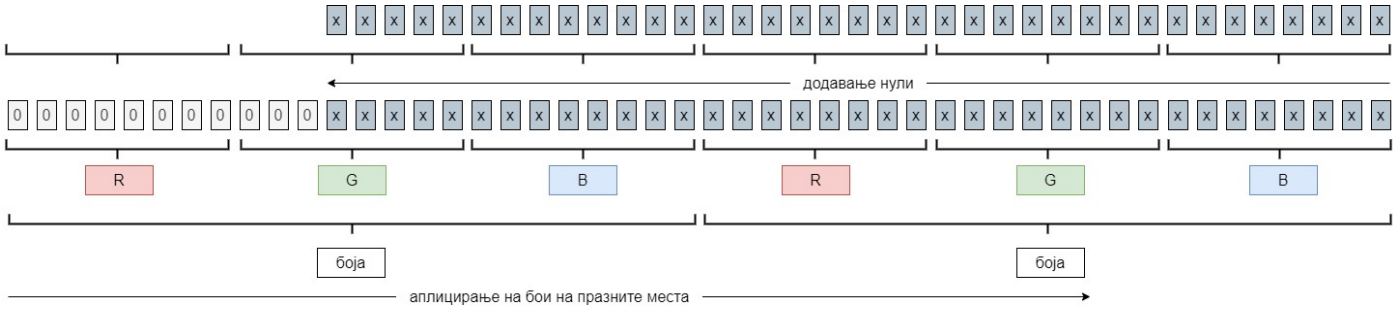
***Чекор 6*:** Како вредност на *G*-параметрот (од *RGB* бојата) се поставува втората децимална вредност од соодветниот кластер (супер-група).

***Чекор 7*:** Како вредност на *B*-параметрот (од *RGB* бојата) се поставува третата децимална вредност од соодветниот кластер (супер-група).

***Чекор 8*:** За следниот невидлив знак во документот, се користи следниот кластер (супер-група) и се повторуваат чекорите од 5 до 8.

Капацитетот на овој метод овозможува пренос на голем број на битови, а со тоа и долги пораки во мали документи и тоа е главната карактеристика со која

што овој тип на метод се одделува од останатите. Секоја појава на невидлив знак, е потенцијален преносител на 24 бита т.е. потенцијален преносител на три знаци од скриената порака.



Слика 3: Алгоритам за обојување на празните места

Недостаток на овој метод е што бојата на невидливиот знак може да е која било зададена боја од *RGB* спектарот. Крајниот корисник не е свесен за бојата на празните места, но доколку се обиде да ја менува содржината на документот, кликнува/селектира некое од обоените празни места и почнува да пишува нова содржина, според карактеристиките на *Microsoft Word*, новата содржина која што ќе биде внесена на овој начин, ќе биде обоена со истата онаа боја на невидливиот знак од каде што започнало пишувањето, па крајниот корисник бргу ќе забележи дека самиот документ содржи нетипични промени во празните места.

Методот претставен во [23] користи четири специјални знаци од софтверот *Microsoft Word*, кои не се видливи за крајниот корисник. Во продолжение се дадени неколку користени симболи во овој метод:

Табела 6: Невидливи знаци кои се користат за вметнување информации помеѓу видливте знаци

име на симбол	Код	кратенка
Right remark	U+200E	*RR*
Left remark	U+200F	*LR*
Zero width joiner	U+200D	*ZWJ*
Zero width non-joiner	U+200C	*ZWNJ*
***Zero width character*	U+200B	*ZWC*

***симболот *ZWC*, се користи во другите методите спомнати подолу

Комбинирањето на четири од овие знаци во определен редослед, резултира со 16 различни комбинации, што значи дека помеѓу секои два знаци, може да се врши пренос на четири бита. Од овде произлегува дека капацитетот на методот директно зависи од бројот на знаци во документот, при тоа што за секој знак врши пренос на четири бита.

Во *Табела 7* е даден пример за тоа, како редоследот на скриените знаци е претходно дефиниран и во зависност од појавата / отсуството на одреден(и) знак(и) се детектира скриена порака од четири бита.

Табела 7: Скриена порака дефинирана со редоследот на невидливите знаци

Right remark	*Left remark*	*Zero width joiner*	*Zero width non-joiner*	скриена порака
x	x	X	x	0000
x	x	X		0001
x	x		x	0010
x	x			0011
x		X	x	0100
x		X		0101
x			x	0110
x				0111
	x	X	x	1000
	x	X		1001
	x		x	1010
	x			1011
		X	x	1100
		X		1101
			x	1110
				1111

Во примерот од *Табела 7*, редоследот на невидливите знаци е дефиниран како „*RR LR ZWJ ZWNJ*“ и постоењето на сите знаци се дефинира како скриена порака 0000, постоењето само на првите три се дефинира како скриена порака 0001, отсуството на сите знаци се дефинира како скриена порака 1111, итн.

Предноста на овој алгоритам е тоа што преносот на скриените битови не влијае врз содржината на документот и не влијае врз форматот и својствата на знаците. Може да се применува за криење на скриени пораки во документи, без разлика каков вид на *Unicode* или *ASCII* знаци се користат.

Методот [24] се базира врз користење на специјалниот невидлив знак *ZWC* од *Табела 6* т.е. негово вметнување до празните места – процес кој минува незабележливо од страна на крајниот корисник, со оглед на тоа дека овој знак не зафаќа никаков простор и не е видлив. Методот претежно ги искористува празните места помеѓу зборовите и реченициите.

Во продолжение се дадени чекорите за вметнување на скриената порака во документот:

***Чекор 1*:** Почнувајќи од десно кон лево, низата од единици и нули која што ја претставува скриената порака, се дели во групи од по два бита. Доколку последната група од битови (онаа на најлевата страна) е група која што има еден бит, на почетокот се додава една бинарни нула, со цел групата да се доведе до големина од два бита.

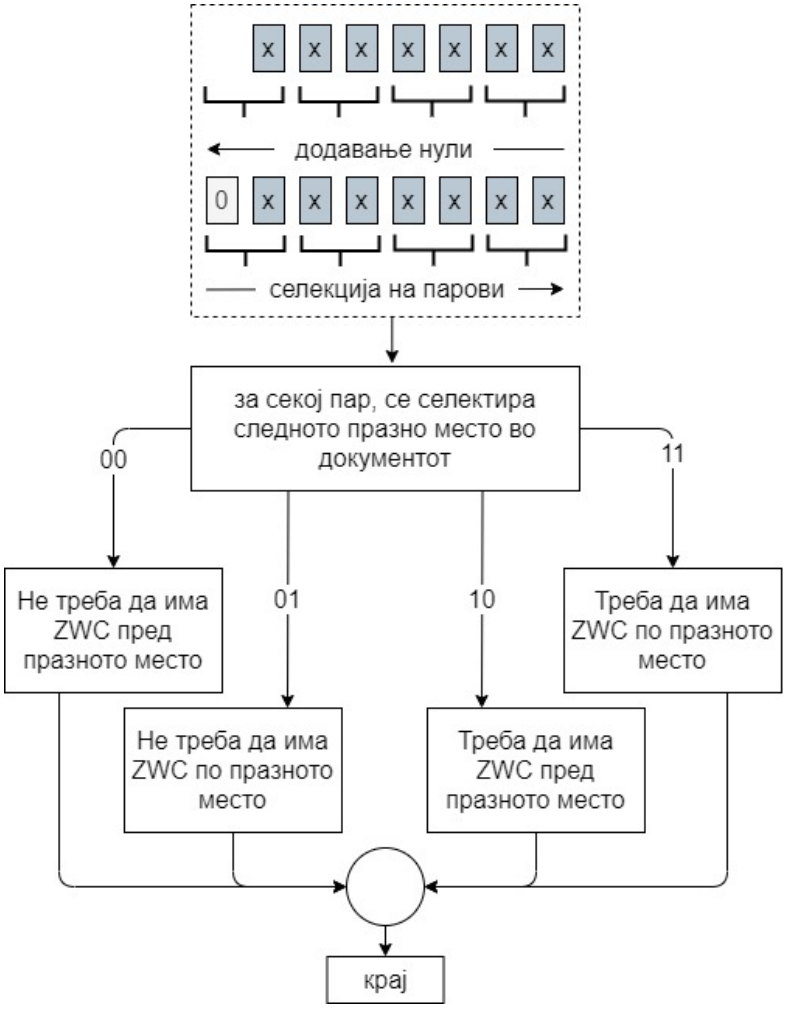
Чекор 2: За првото празно место во документот се користи првата група од скриената порака.

***Чекор 3*:** Се врши проверка на вредноста на групата и во зависност од нејзината вредност, се вршат соодветни промени:

- доколку групата = 00, тогаш пред празното место не треба да има *ZWC*;
- доколку групата = 01, тогаш по празното место не треба да има *ZWC*;
- доколку групата = 10, тогаш пред празното место треба да има *ZWC*;
- доколку групата = 11, тогаш по празното место треба да има *ZWC*.

***Чекор 4*:** За следниот невидлив знак во документот, се користи следната група и се повторува чекор 3.

Истиот е даден и на *Слика 4 и Слика 3* подолу, од каде што јасно се гледа како се врши групирањето на битовите од скриената порака во парови и како врз основа на добиените парови, се врши манипулации со празните места т.е. се дефинира дали ќе се вгнезди *ZWC* карактер пред и по празното место.



Слика 4: Алгоритам за вметнување на ZWC карактер до празните места

Капацитетот на овој метод зависи од бројот на празни места во документот, на начин што едно празно место е преносител на два бита од скриената порака.

Методот [25] има безбедност на повеќе нивоа, бидејќи процесот на вметнување на скриената порака има повеќе фази. Овој метод го користи специјалниот невидлив знак *ZWC*, како и повеќе видови на празни места [26] дадени во *Табела 8*.

Процесот започнува така што, откако скриената порака се претвора во бинарни единици и бинарни нули, се применуваат чекорите кои што се дадени во продолжение.

Табела 8: Типови на празни места и „zero-width“ празни места во Unicode

име на симбол	код	пример	ширина на карактер
Space	U+0020	foo bar	зависи од фонтот, обично ¼ em, најчесто се прилагодува
No-break space	U+00A0	foo bar	зависи од фонтот, обично ¼ em, најчесто не се прилагодува
Ogham space mark	U+1680	foo bar	недефинирано; често не е празно место туку линија
Mongolian vowel separator	U+180E	foobar	0
En quad	U+2000	foo bar	1 en (= 1/2 em)
Em quad	U+2001	foo bar	1 em (евентуално, висината на фонтот)
En space (nut)	U+2002	foo bar	1 en (= 1/2 em)
Em space (mutton)	U+2003	foo bar	1 em
Three-per-em space (thick space)	U+2004	foo bar	1/3 em
Four-per-em space (mid space)	U+2005	foo bar	1/4 em
Six-per-em space	U+2006	foo bar	1/6 em
Figure space	U+2007	foo bar	„табуларна ширина“, ширината на броевите
Punctuation space	U+2008	foo bar	ширината на точка „.“
Thin space	U+2009	foo bar	1/5 em (или понекогаш 1/6 em)
Hair space	U+200A	foo bar	малку помало од „Thin space“
Zero width character / Zero width space	U+200B	foobar	0
Narrow no-break space	U+202F	foo bar	Малку помало од „No-break space“ (или „Space“), обично ширината на „Thin space“ или „Mid space“
Medium mathematical space	U+205F	foo bar	4/18 em
Ideographic space	U+3000	foo bar	ширината на „ideographic (CJK) characters“ [27]
Zero width no-break space	U+FEFF	foobar	0

***Чекор 1*:** Се врши пермутација на бинарната порака (менување на редоследот на знаците) со користење на таен клуч.

***Чекор 2*:** Се врши инверзија, со што бинарните единици стануваат бинарни нули, а бинарните нули стануваат бинарни единици.

***Чекор 3*:** Почнувајќи од десно кон лево, низата од единици и нули која што ја претставува скриената порака, се дели во групи од по четири бита. Доколку последната група од битови (онаа на најлевата страна) е група која што има помалку од четири бита, на почетокот се додаваат бинарни нули, со цел групата да се доведе до големина од четири бита.

Чекор 4: Се врши компресија, користејќи ја *Табела 9* – процес што е опишан подолу. Овој чекор резултира со нова бинарна порака, која што е два пати помала од пораката добиена во чекор 3.

***Чекор 5*:** Се врши проверка на капацитетот т.е. дали оригиналниот документ има доволен број на празни места. Едно празно место врши пренос на четири бита, па доколку бројот на празни места во оригиналниот документ >= (бројот на битови од компресираната порака (од чекор 4)) / 4, тогаш може да се продолжи понатаму.

***Чекор 6*:** Се врши селекција на случајни парови на знаци од *Табела 8,* кои се користат како празни места – процес што е опишан подолу.

Чекор 7: Се врши вгнездување на скриената порака – процес што е опишан подолу. За првото празно место во документот се користат првите два / четири бита од компресирната скриена порака (од чекор 4), во зависност од класата на празното место.

***Чекор 8*:** За следното празно место во документот, се користи следната група од два / четири бита, во зависност од класата на празното место и се повторува чекор 7.

Компресија: Како влезен параметар во овој процес е бинарна порака која што е поделена на групи од по четири бита и се пристапува така што пораката се разгледува група по група. Сите можни комбинации на една група од четири бита се дадени во првата колона на *Табела 9*. Секоја комбинација се заменува со соодветната мапира вредност од втората колона и се означува како дел од соодветната група од третата колона.

Табела 9: Мапирање при компресија

комбинација	мапирана вредност	група
0000	00	*G1*
0001	01	*G1*
0010	10	*G1*
0011	11	*G1*
0100	00	*G2*
0101	01	*G2*
0110	10	*G2*
0111	11	*G2*
1000	00	*G3*
1001	01	*G3*
1010	10	*G3*
1011	11	*G3*
1100	00	*G4*
1101	01	*G4*
1110	10	*G4*
1111	11	*G4*

Комбинациите може да се означат како дел од четири групи *G1*, *G2*, *G3* или *G4* така што групите се поделени на начин што сите четири комбинации кои одговараат на одредена група, започуваат со исти два бита, т.е.:

- четирите комбинации од *G1* започнуваат со 00;
- четирите комбинации од *G2* започнуваат со 01;
- четирите комбинации од *G3* започнуваат со 10;
- четирите комбинации од *G4* започнуваат со 11.

На овој начин, ознаката за тоа која комбинација одговара на која група и чувањето на редоследот на групите во посебна мапа, овозможува комбинациите од првата колона да се заменат со нивните мапирани вредности од втората колона, со напомена дека при процесот на декодирање на скриената порака, редоследот на групите од третата колона треба да е достапен за соодветна декомпресија на скриената бинарна порака.

Процесот на компресија всушност како влезен параметар прима бинарна порака поделена на групи од по четири бита, а како излезен параметар ги елиминира првите два бита од секоја група од влезниот параметар, со што како излезен параметар се добива двојно помала скриена порака.

Селекција на случајни парови: Еден од начините за криење на битови во документите при процесот на вгнездување (опишан подолу) е промена на регуларното празно место со друг вид на празно место / или со комбинација од празни места, што всушност поминува незабележително од страна на крајниот корисник. Со оглед на тоа дека постојат повеќе знаци кои може да се користат како замена за регуларното празно место (дадени во *Табела 8*), во овој чекор се врши селекција на алтернативни знаци (или парови од знаци) од дадените можности во табелата. Празните места во *Microsoft Word* документите се поделени во две класи:

A. празни места помеѓу зборовите и празни места помеѓу реченициите;
B. празни места на крајот на линиите и празни места помеѓу параграфите.

Замената на празните места се врши според нивните класи т.е. празните места од класа А се заменуваат со едно мапирање, додека празните места од класа В се заменуваат со друго мапирање. Еден пример на селекција на мапирање за класа А е даден во *Табела 10* и за класа В е даден во *Табела 11*.

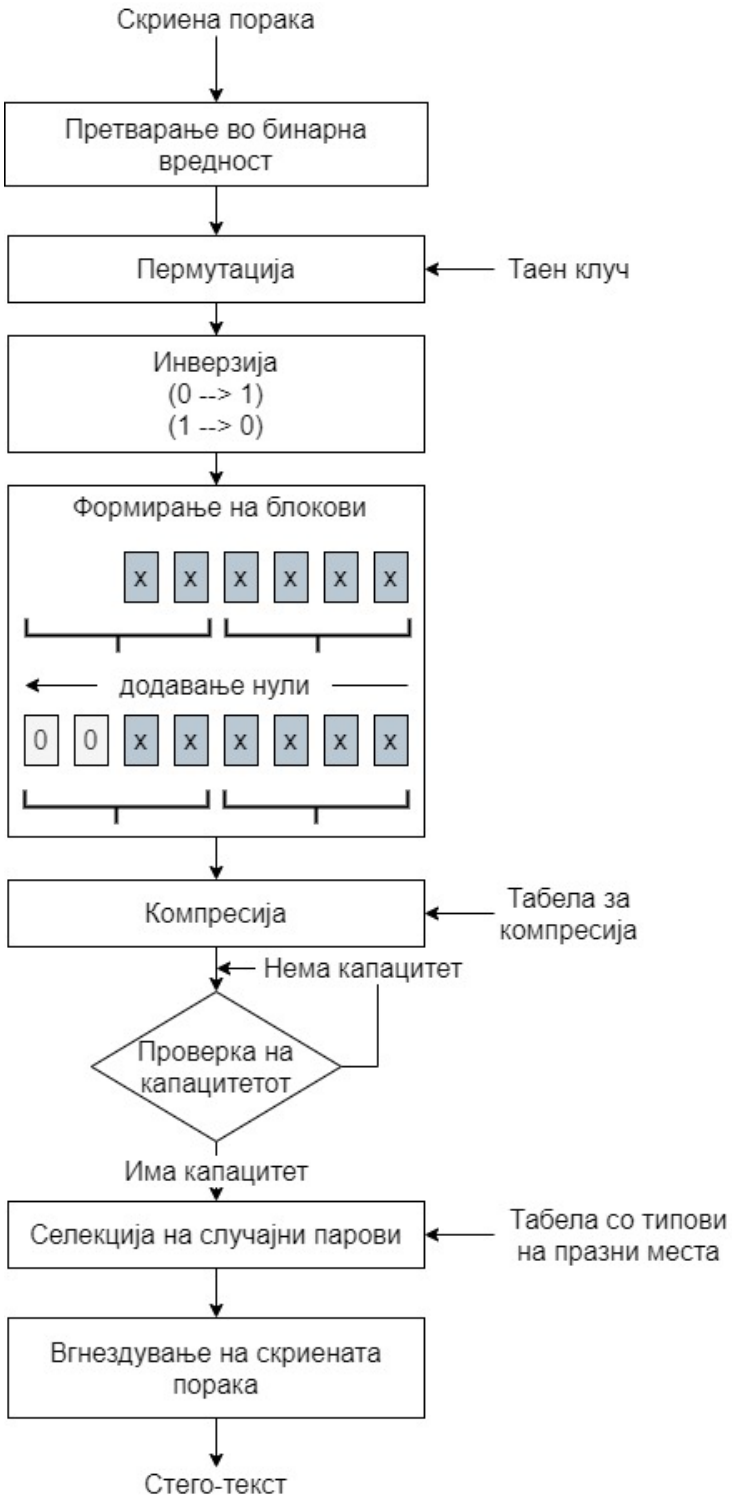
Табела 10: Мапирање на празни места од класа А

комбинација	секвенца
Normal	00
Thin + Normal	01
Six-Per-Em + Normal	10
Hair + Normal	11

Табела 11: Мапирање на празни места од класа В

карактер	секвенца
Hair	00
Six-Per-Em	01
Punctuation	10
Thin	11

До празните места помеѓу зборовите и помеѓу реченициите се додаваат знаците од *Табела 10*. Тие се избрани бидејќи се всушност знаците со најмала ширина од *Табела 8* и со тоа имаат најмало влијание врз визуелниот ефект на оригиналниот документ. Празните места за премин во нова линија и премин во нов параграф, се директно заменуваат со знаците од *Табела 11*.



Слика 5: Алгоритам за вметнување на различни типови на празни места

Целокупниот процес на овој алгоритам, од скриената порака како влезен параметар, па се до добивањето на документот во кој е вгнездена скриената порака, е прикажан со шемата на *Слика 5*.

Вгнездување на скриената порака: компресираната бинарна порака се вгнездува во празните места, во зависност од класата на празното место. Доколку празното место е од класа А, тогаш тоа е преносител на четири бита од компресираната порака. Првите два бита се вгнездуваат со замена на празното место со соодветната комбинација од *Табела 10* во зависност од тоа на која секвенца битовите одговараат. Другите два бита се вгнездуваат во вметнување на *ZWC* карактерот во празното место, на следниот начин:

- доколку секвенцата е 00, тогаш не се вметнува *ZWC*;
- доколку секвенцата е 01, тогаш *ZWC* се вметнува по комбинацијата од празните места претходно добиени со мапирање од *Табела 10*;
- доколку секвенцата е 10, тогаш *ZWC* се вметнува пред комбинацијата од празните места претходно добиени со мапирање од *Табела 10*;
- доколку секвенцата е 11, тогаш *ZWC* се вметнува пред и по комбинацијата од празните места претходно добиени со мапирање од *Табела 10*.

Доколку пак празното место е од класа В, тогаш тоа празно место е преносител на два бита од компресираната порака т.е. тие се вгнездуваат со замена на празното место со соодветниот знак од *Табела 11*, во зависност од тоа на која секвенца битовите одговараат.

2.2..7 Метод базиран на типови на фонт

Методот [28] врши вгнездување на скриената порака преку искористување на можноста за промена на типот на фонтот на текстот. Софтверот *Microsoft Word* содржи значително големо множество од достапни типови на фонтови, при што поради нивната бројност и разновидност, дел од нив имаат многу слични карактеристики, т.е. нивните стилови се многу слични до степен каде што имаат само некои минимални разлики.

Овој метод ја искористува токму таа сличност помеѓу фонтовите, со што се менуваат типовите на фонтовите на буквите со некои од типовите кои имаат минимални разлики. Дополнително, бидејќи истражувањето при развојот на методот покажало дека сличностите помеѓу фонтовите се поголеми кај големите букви, со цел промените да не бидат лесно детектирани од страна на обичниот корисник, истите се применуваат само врз големите букви во документот.

Друга разлика од досега разгледаните техники е тоа што методот [28] не се базира на криење на бинарна порака од единици и нули (кои подоцна се претвораат во формат читлив за човекот), напротив, тој директно поддржува вгнездување само на буквите од англиската азбука и на знакот за празни место. Со тоа методот нема можност за криење на кој било знак, како: интерпункциски знаци, специјални знаци, разлика помеѓу големи и мали букви, итн.

За типот на фонтот кој се користи во оригиналниот документ, се наоѓаат три други фонтови кои се слични на оригиналниот и кои се користат за соодветна замена кај големите букви. Како пример во *Табела 12* се дадени едни од најкористените (15) типови на фонтови и за секој од нив се дадени по три слични фонта. Самата табела може да служи како референца доколку кои било од дадените типови на фонтови се јавува како фонт во оригиналниот документ, т.е. останатите три фонтови во соодветниот ред, може да се искористат како алтернативи за негово мапирање.

Табела 12: Фонт користен во оригиналниот документ и негови слични фонтови за мапирање

фонт во оригинален документ	алтернатива 1	алтернатива 2	алтернатива 3
Arial	*Arial Unicode MS*	*Geo_Arial*	*Microsoft Sans Serif*
Book Antiqua	*Palatino Linotype*	*Antiqua*	*Caudex*
Candara	*Khmer UI*	*Ebrima*	*Microsoft New Tai Lue*
Century	*Century751 BT*	*CenturyOldStyle*	*CenturyExpd BT*
Calibri	*Gisha*	*Leelawadee*	*Liberation Serif*
Cambria	*Proforma*	*EideticNeoRegular*	*Liberation Serif*
Comic Sans	*SF Toontime*	*Komika Text*	*SF Arch Rival Extended*
Times New Roman	*Tinos*	*Liberation Serif*	*Thorndale*
Helvetica	*Arimo*	*Geo_Arial*	*Arial-Relcom*
Courier New	*Courier New CE*	*TiredOfCourier*	*TiredOfCourierThin*
Verdana	*Tahoma*	*MS Reference Sans Serif*	*Lato*
Perpetua	*ChanticleerRoman*	*Centaur*	*CaslonOldFace BT*
Lucida Sans	*Lucida Sans Unicode*	*Segoe UI*	*Lucida Sans Typewriter*
Thorndale	*Times New Roman*	*Liberation Serif*	*Tinos*
Franklin Gothic Book	*Ebrima*	*Corbel*	*Trebuchet MS*

При кодирањето, секој симбол од скриената порака може да се претстави преку три типови на фонт т.е. 27 знаци (буквите од англиската азбука и празното

место) може да бидат сокриени во три големи букви во оригиналниот документ, користејќи ги трите достапни алтернативи за замена на фонтот. На пример, во табелата е дадено дека слични фонтови на *Century* се:

Century = {Century751 BT, CenturyOldStyle, CenturyExpdBT}

Па така, доколку симболот кој треба да се сокрие е претставен преку групата (1, 1, 1) тогаш следните три големи букви ќе се заменат и наместо да се користи оригиналниот фонт, кај истите ќе се користи (алтернатива 1, алтернатива 1, алтернатива 1) соодветно. Доколку симболот кој треба да се сокрие е претставен преку групата (1, 2, 2) ќе се користи (алтернатива 1, алтернатива 2, алтернатива 2) соодветно, за групата (3, 1, 2) ќе се користи (алтернатива 3, алтернатива 2, алтернатива 1) итн.

Вгнездувањето на пораката започнува од појавата на првата голема буква во оригиналниот документ и за секој знак од скриената порака (буквите и празното место) се користат следните три големи букви во оригиналниот документ. Од овде произлегува дека големите букви се всушност ентитетите кои го вршат преносот на скриената порака и капацитетот за пренос зависи директно од нивната бројност во документот. Скриената порака завршува кога ќе биде препознаена групата (0, 0, 0) т.е. три последователни големи букви во оригиналниот документ, го имаат оригиналниот фонт.

Како што е споменато погоре, со цел успешно да се изврши вгнездувањето на симболите од скриената порака во големите букви на оригиналниот документ, секој симбол од скриената порака треба да биде претставен преку група од три вредности. Ова мапирање го дефинира лицето кое го врши вгнездувањето и како такво не секогаш мора да биде исто. Еден пример на вакво мапирање е даден во *Табела 13*.

Двете табели разгледани повторно, даваат појасна слика за тоа на кој начин функционира опишаниот метод: вредностите во колоните во табелата со кодови всушност може да се разгледуваат како бројот на алтернативниот фонт од *Табела 12* кој што треба да го замени оригиналниот фонт на местото на соодветниот симбол од *Табела 13.*

Табела 13: Табела на кодови за буквите од англиската азбука и празното место

Карактер	симбол 1	симбол 2	симбол 3
A	1	1	1
B	1	1	2
C	1	1	3
D	1	2	1
E	1	2	2
F	1	2	3
G	1	3	1
H	1	3	2
I	1	3	3
J	2	1	1
K	2	1	2
L	2	1	3
M	2	2	1
N	2	2	2
O	2	2	3
P	2	3	1
Q	2	3	2
R	2	3	3
S	3	1	1
T	3	1	2
U	3	1	3
V	3	2	1
W	3	2	2
X	3	2	3
Y	3	3	1
Z	3	3	2
празно место	3	3	3

2.2..8 Метод базиран на повеќејазични *Unicode* знаци

Методот базиран на повеќејазични *Unicode* знаци [29] го користи фактот што дел од знаците (буквите) од англиската азбука се појавуваат како знаци и во други азбуки, но со различен *Unicode*. Дел од нив, имаат и различен стил (закосени краеви и сл.), па само 13 од знаците се погодни за користење во овој метод.

Методот е наменет само за документи чија што оригинална содржина се знаци од основниот *Unicode* систем (*Base Multilingual Plane*) чии што вредности се во рангот *U+0000 – U+FFFF* и се базира на предефинирана табела на *ASCII / Unicode* мапирање. Во *Табела 14* е прикажан начинот на кој секои два бита од скриената бинарна порака, може да се вгнездат со користење на една буква (доколку буквата е една од оние 13 знаци кои се погодни за користење во овој метод). Вгнездувањето се врши на начин што оригиналната буква се заменува со знакот кој ја има соодветната *Unicode* вредност, во зависност од вредноста на скриената порака. Со оглед на тоа дека знаците во табелата се изберени на начин што сите имаат многу големи сличности едни со други, крајниот корисник не е во можност да ја забележи промената.

Табела 14: Букви од англиската азбука кои имаат соодветни Unicode знаци

симбол	*ASCII*	*Unicode*		
	тајна порака 00	тајна порака 01	тајна порака 10	тајна порака 11
A	0041	0391	0410	13AA
B	0042	0392	0412	0181
E	0045	0395	0415	13AC
G	0047	050C	12C0	13B6
H	0048	0397	041D	13BB
I	0049	0399	04C0	0406
M	004D	039C	041C	216F
O	004F	039F	041E	0555
P	0050	0420	03A1	01A4
S	0053	0405	054F	13DA
T	0054	0422	03A4	01AC
j	006A	0458	03F3	029D
o	006F	03BF	1D0F	043E

Како што може да се забележи од табелата, доколку следните два бита од скриената порака се 00, знакот останува ист т.е. не се менува. Замената со новиот знак се врши само доколку следните два бита од скриената порака имаат една од вредностите 01, 10 или 11.

Дополнително, пред да започне процесот на вгнездување на скриената порака, се детектира должината на истата и се вметнува во првите 16 бита од вгнездената порака во оригиналниот документ.

Како ентитети кои го вршат преносот, се јавуваат само дадените 13 знаци, па капацитетот на методот директно зависи од бројот на појавувања на истите во документот.

2.3 Стеганографски лингвистички методи

Лингвистичките методи за стеганографија вршат манипулирање во лексичките, синтаксичките или семантичките својства на текстот, при што значењето на текстот е зачувано колку што е можно повеќе. Главни поткатегории на лингвистичките методи се синтаксичките методи и семантичките методи.

Кај синтаксичките методи, скриената порака се вгнездува со самата синтаксичка структура. Понекогаш вгнездувањето се постигнува во менување на дикцијата и структурата на текстот, без притоа да се изврши значителна промена на суштинското значење на оригиналниот текст. На пример, промените може да се вршат со користење на интерпункциските знаци, со оглед на тоа дека тие често може да се додаваат на места каде што всушност не се ни потребни, а доколку се додадат на место каде што граматички тоа не е правилно, истите имаат незначително влијание на значењето на текстот.

Од друга страна пак, кај семантичките методи, скриената порака се вгнездува со менување на самите зборови и карактеристично за овој тип на методи е тоа дека во секој документ не може да се скрие која било порака. Додека во методите базирани на формати главен показател за тоа дали може да се вгнезди една порака или не, е само капацитетот на методот (изразен преку фреквенцијата на ентитетите за пренос – кои се различни од метод до метод), кај оваа категорија на лингвистички методи, методологијата за тоа дали скриената порака може да се вгнезди или не, е покомплексна и зависи не само од фреквенцијата на ентитетите, туку директно зависи и од самата содржина во документот. Од овде произлегува дека понекогаш самите документи кои ја пренесуваат скриената порака често се наменски креирани за пренос на истата таа порака, за разлика од методите базирани на формати кои може да се применат на кој било документ кои го има соодветниот капацитет за пренос.

2.3..1 Метод за следење на промени на зборовите

Еден пример на семантички метод е кога за време на процесот на вгнездување на скриената порака во оригиналниот документ, се врши промена на оригиналните зборови во документот со нивни соодветни – погрешно спелувани зборови. Методот [30] се базира на бројот на појавувања на секој од зборовите во документот и ова пребројување претставува почетна точка во процесот на вгнездување.

На пример, како текст во оригиналниот документ нека е даден:

The quick brown fox jumped over the lazy dog.

Врз база на фреквенцијата на зборовите во документот, се креира дрво на *Huffman* [31] со бинарни вредности. Резултатот од овој процес е тоа што по креирањето на дрвото, секој збор добива своја локација (јазол) во дрвото и со тоа е мапиран во свој бинарен код (кој се добива почнувајќи од коренот на дрвото до соодветниот јазол). Дополнително, за секој збор се наоѓа негова соодветна замена т.е. погрешно спелуван збор кој ќе биде негова алтернатива во случај да треба да се врши менување на зборот.

Табела 15: Бинарни кодови добиени со креирање на дрво на Huffman

оригинален збор	погрешно спелуван збор	бинарен код
The	*Teh*	11
Quick	*Qwik*	101
Brown	*Borwn*	10010
Fox	*Foxx*	0011
Jumped	*Jumpped*	000010
Over	*Oevr*	0001110
Lazy	*Lazzy*	100010
Dog	*Dag*	01

Во *Табела 15* всушност е претставено како секој збор (прва колона) се мапира во соодветен бинарен код (трета колона), во случај кога истиот е заменет со соодветен погрешно спелуван збор (втора колона).

Соединувањето на бинарните кодови на погрешно спелуваните зборови резултира со подолга бинарна порака, која што може да се третира како *ASCII* код, кој потоа може да се декодира во скриена порака. На пример, со соединување на бинарните кодови за „*brown*“, „*over*“ и „*dog*“ се добива скриената порака “Hi”:

10010 0001110 01 → 1001000 0111001 → 0x72 0x105 → Hi

Резултатот по вгнездувањето т.е. по извршеното менување на зборовите, во кој е вгнездена скриената порака „Hi“ е даден како:

The quick borwn fox jumped oevr the lazy dag.

Следењето на промените на зборовите би било претставено како:

The quick ~~borwn~~brown fox jumped ~~oevr~~over the lazy ~~dag~~dog.

Како што може да се забележи од самиот пристап, претходната констатација за тоа дека скриената порака често зависи директно од самата содржина на документот се потврдува дури и од дадениот едноставен пример т.е. со оглед на тоа дека дрвото на *Huffman* се креира врз основа на фреквенцијата на зборовите, доколку потребните *ASCII* вредности (за скриената порака) не можат да бидат креирани, потребно е наменски да се менува содржината на оригиналниот документ. Целта е да преку додавање или отстранување на одредени зборови, дрвото на *Huffman* да се доведе во состојба од која што ќе може да се користат бинарните кодови, при формирањето на *ASCII* вредностите за скриената порака.

Главен недостаток на опишаниот метод е тоа што кога крајниот корисник во гледа документот со вгнездената скриена порака, очигледно е дека нешто не е во ред со истиот, со оглед на тоа дека содржи поголем број на погрешно спелувани зборови што понекогаш може да доведе и до менување на самата содржина на оригиналниот документ.

Алтернатива на опишаниот метод е наместо менување на редоследот на знаците во зборовите (погрешно спелувани зборови), да се менува големината на буквите во зборовите, на места каде што од граматичка гледна точка, тоа не треба да се случува. На пример, ако зборот се наоѓа на средината во

реченицата, тогаш не е очекувано тој да започнува со голема буква т.е. доколку тоа е случај, истиот може да се третира како граматичка грешка во оригиналниот текст и истовремено да претставува ентитет кој го врши преносот на скриените битови. И покрај тоа што предност на оваа алтернатива е тоа што не го менува значењето на оригиналната содржина, сепак може лесно да се детектира од страна на крајниот корисник, кој не би требало да биде свесен за постоењето на скриената порака во документот.

2.3..2 Методи за мапирање на зборовите

Кај постоечкиот метод за мапирање на зборовите [32], пред да се пристапи кон процесот на вгнездување на скриената порака, се врши шифрирање на истата со цел додавање на дополнително ниво на безбедност. Па во случај алгоритамот да е препознаен и пробиен од страна на трето лице кое што не би требало да ја прочита пораката, при самото декодирање на стеганографскиот метод ќе се добие порака чија што содржина сè уште нема да има смисла, поради тоа што истата е дополнително шифрирана.

Шифрирањето се врши на начин што секој од знаците на скриената порака се претвора во неговиот *ASCII* код и потоа користејќи ја табелата од *Слика 6* секој добиен код се заменува со друг, т.е. со соодветниот *ASCII* код од сликата. Оваа табела на мапирање се нарекува табела на тајни стеганографски кодови за вгнездување.

По извршеното шифрирање, се добива низа од нови *ASCII* кодови кои се претвораат во нивните соодветни бинарни вредности и добиената бинарна порака по ова шифрирање, е онаа која што се вгнездува во оригиналниот документ. Вгнездувањето се врши на начин што се менуваат вредностите на зборовите „*a*“ и „*an*“ еден со друг, според *Табела 16*.

Во граматиката на англискиот јазик, граматички правилно е кога пред зборовите што започнуваат со согласка да стои „*a*“, додека пред зборовите што започнуваат со самогласка да стои „*an*“ и овие комбинации се мапираат во битовите 00 / 11 соодветно. Промената на граматиката е предизвикана кога се јавува потреба за пренос на битовите 01 и 10.

Со анализа на наведениот метод се доаѓа до заклучок дека како ентитети за пренос на скриените битови се зборовите кои пред себе содржат еден од

зборовите „*a*“ или „*an*“, и двата зборови заедно во комбинација претставуваат ентитет за пренос. Капацитетот на методот зависи од постоењето на овие комбинации, со тоа што секоја комбинација од два збора пренесува два бита.

Содржината на оригиналниот документ кај овој метод не е толку многу строго поврзана со скриената порака која ја содржи, т.е. доколку има доволно ентитети преносители, истите ќе се прилагодат според скриената порака. Сепак, методот може да биде забележлив од крајниот корисник, со оглед на тоа дека колку е поголема скриената порака, документот ќе содржи повеќе граматички грешки.

Табела 16: Техника за мапирање на зборовите

Зборови		Битови
A	Согласка	00
An	Самогласка	11
A	Самогласка	10
An	Согласка	01

ASCII	SSCE	ASCII	SSCE	ASCII	SSCE	ASCII	SSCE	ASCII	SSCE	ASCII	SSCE	ASCII	SSCE	ASCII	SSCE	ASCII	SSCE	ASCII	SSCE
10	1	1	26	2	52	3	78	4	104	5	130	6	156	7	181	8	206	9	231
20	2	11	27	12	53	13	79	14	105	15	131	16	157	17	182	18	207	19	232
30	3	21	28	22	54	23	80	24	106	25	132	26	158	27	183	28	208	29	233
40	4	31	29	32	55	33	81	34	107	35	133	36	159	37	184	38	209	39	234
50	5	41	30	42	56	43	82	44	108	45	134	46	160	47	185	48	210	49	235
60	6	51	31	52	57	53	83	54	109	55	135	56	161	57	186	58	211	59	236
70	7	61	32	62	58	63	84	64	110	65	136	66	162	67	187	68	212	69	237
80	8	71	33	72	59	73	85	74	111	75	137	76	163	77	188	78	213	79	238
90	9	81	34	82	60	83	86	84	112	85	138	86	164	87	189	88	214	89	239
100	10	91	35	92	61	93	87	94	113	95	139	96	165	97	190	98	215	99	240
110	11	101	36	102	62	103	88	104	114	105	140	106	166	107	191	108	216	109	241
120	12	111	37	112	63	113	89	114	115	115	141	116	167	117	192	118	217	119	242
130	13	121	38	122	64	123	90	124	116	125	142	126	168	127	193	128	218	129	243
140	14	131	39	132	65	133	91	134	117	135	143	136	169	137	194	138	219	139	244
150	15	141	40	142	66	143	92	144	118	145	144	146	170	147	195	148	220	149	245
160	16	151	41	152	67	153	93	154	119	155	145	156	171	157	196	158	221	159	246
170	17	161	42	162	68	163	94	164	120	165	146	166	172	167	197	168	222	169	247
180	18	171	43	172	69	173	95	174	121	175	147	176	173	177	198	178	223	179	248
190	19	181	44	182	70	183	96	184	122	185	148	186	174	187	199	188	224	189	249
200	20	191	45	192	71	193	97	194	123	195	149	196	175	197	200	198	225	199	250
210	21	201	46	202	72	203	98	204	124	205	150	206	176	207	201	208	226	209	251
220	22	211	47	212	73	213	99	214	125	215	151	216	177	217	202	218	227	219	252
230	23	221	48	222	74	223	100	224	126	225	152	226	178	227	203	228	228	229	253
240	24	231	49	232	75	233	101	234	127	235	153	236	179	237	204	238	229	239	254
250	25	241	50	242	76	243	102	244	128	245	154	246	180	247	205	248	230	249	255
		251	51	252	77	253	103	254	129	255	155								

Слика 6: Мапирање на тајни стеганографски кодови за вгнездување

Погоре опишаниот метод се развива како идеја на истите автори кои го развиле методот [33], каде што криењето на скриената порака се врши преку вметнување на празни места помеѓу соседни зборови при што двата соседни зборови имаат иста парност т.е. и двата имаат или парен или непарен број на знаци. Ентитети за пренос се комбинацијата од зборови со иста парност и секоја комбинација е преносител на два скриени битови.

2.3..3 Метод базиран на менување на спелувањето на зборовите

Англискиот јазик како еден на најкористените јазици во светот, има две говорни подрачја: англиски јазик кој се користи во Обединетото Кралство и англиски јазик кој се користи во Соединетите Американски Држави и насекаде низ светот. Двете говорни подрачја во основа се исти, но сепак имаат некои меѓусебни разлики како што се различното спелување на некои од зборовите. Овој метод [34] се базира токму на тие разлики и во зависност од тоа кое спелување на зборот се користи, се извршува пренос на бинарна единица или бинарна нула.

Разликата од претходниот метод со спелување на зборовите е тоа што претходно стануваше збор за погрешно спелување на зборови кое што лесно се детектира, но во овој случај зборовите не се спелуваат погрешно, туку напротив само се користи различно говорно подрачје кое што во суштина е сè уште граматички точно.

Се пристапува кон креирање на табела (како на пример *Табела 17*) при што се дадени зборови кои имаат различно спелување во двете говорни подрачја и токму овие зборови претставуваат ентитети за пренос на скриената порака.

Секое појавување на зборот спелуван како во САД се интерпретира како бинарна нула од скриената порака и секое појавување на зборот спелуван како во Обединетото Кралство се интерпретира како бинарна единица од скриената порака.

Капацитетот на методот зависи од фреквенцијата на појавата на избраните зборови (од табелата која што ќе се користи), па со оглед на тоа дека секој збор пренесува еден бит и дека бројот на потенцијални зборови кои може да се додадат во табелата е ограничен, самиот капацитет и не е толку многу голем т.е. методот не е погоден за пренос на долги пораки.

Алтернатива на опишаниот метод е да се задржи истото говорно подрачје низ целиот оригинален документ и на сличен начин да се креира табела со синоними. Со оглед на тоа дека во англискиот јазик има голем број на синоними, зборовите можат да се користат како ентитети за пренос на еден или два бита. Како на пример: појавата на „*big*“ или „*large*“ може да се дефинира како пренос на 0 или 1 соодветно; појавата на „*propensity*“, „*predilection*“, „*penchant*“ или „*proclivity*“ може да се дефинира како пренос на 00, 01, 10 или 11 соодветно, бидејќи сите тие се синоними еден со друг.

Користењето на табелата со синоними е сепак процес кој треба внимателно да се следи, со оглед на тоа дека понекогаш користењето на синоними може да даде поразлично значење на текстот од оригиналниот документ.

Табела 17: Различно спелување на зборови во САД и Обединетото Кралство

Спелување во САД	Спелување во Обединетото Кралство
Favorite	Favourite
Criticize	Criticise
Fulfill	Fulfil
Center	Centre
Dialog	Dialogue
Medieval	Mediaeval
Check	Cheque
Defense	Defence
Tire	Tyre

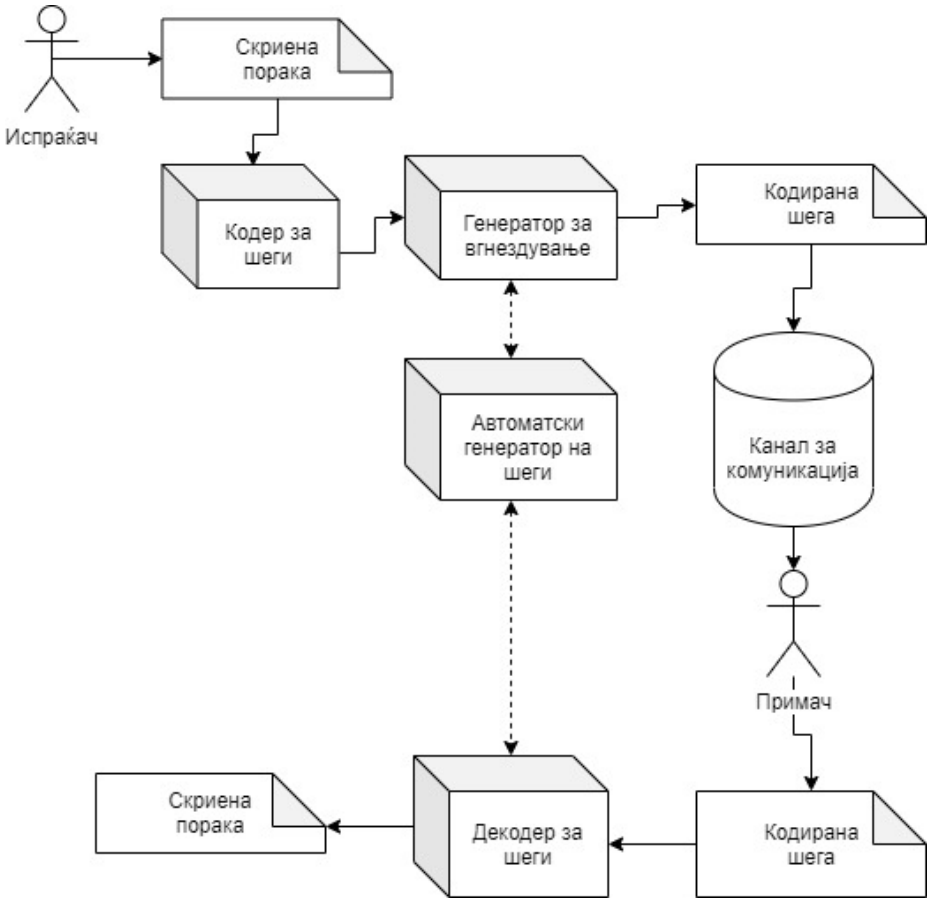
2.4 Стеганографски методи базирани на случајно и статистичко генерирање

Во методите за стеганографија со случајно и статистичко генерирање, се врши генерирање на нов текст кој се обидува да симулира одредени својства на нормален текст, вообичаено со приближување на некои произволни статистички дистрибуции кои се добиени од текстовите во секојдневието. Оваа категорија на методи се разликуваат од претходните две категории, во тоа што тие немаат однапред определено ентитет за пренос на скриената порака, како што во досега опишаните методи тоа најчесто беа знаците / зборовите / различни комбинации

на ентитети. Кај овој тип на методи, преносителот на пораката се генерира автоматски.

2.4..1 Метод базиран на автоматски генератор на шеги

Користењето на кратки шеги како преносители на скриени пораки е идеја која е многу повеќе прифатлива, во смисла дека многу полесно би поминала кај крајниот корисник. Понекогаш и самите оригинални шеги немаат точна граматичка форма, немаат точно спелување, па дури некогаш немаат ни смисла, па од овде потекнува идејата за искористување на овие карактеристики и примена на техники кои би можеле да добијат резултат кој во нормални документи би привлекол посебно внимание од крајниот корисник, но истиот во рамките на шега би бил прифатлив.



Слика 7: Илустрација на интеракцијата на различните модули при генерирање на шегите и како резултатите од индивидуалните модули се користат при конвертирање на комуникацијата помеѓу двете страни

Шемата на методот базиран на автоматски генератор на шеги [35] е дадена на *Слика 7* каде што се прикажани сите модули кои се користат при процесот на испраќање на скриената порака од испраќачот до примачот. Самиот генератор на шеги (врз кој се базира целиот метод) е даден како централен модул во шемата и истиот генерира „оригинална шега“ – текст кој не содржи скриена порака. Вакви типови на генератори се предложени во [36] [37] и [38]. Алгоритамот според кој се врши вгнездувањето на скриената порака (во рамките на генераторот за вгнездување) се дефинира и применува во кодерот и декодерот на шеги.

Конкретен пример за тоа како изгледа резултатот од модулите е даден во продолжение, каде што се претставени етапите на промени на една дадена порака.

Шега генерирана од автоматскиот генератор на шеги:

- Where do milk shakes come from? - Nervous cows!

Вгнездување на скриената порака, во генераторот за вгнездување:

- Where do milk shakes come from? ~~Nervous~~ Shaking cows!

Кодирана шега која ја содржи скриената порака и е доставена до примачот:

- Where do milk shakes come from? Shaking cows!

Секоја генерирана шега има т.н. клучен збор кој претставува ентитет за пренос на скриената порака и тој се менува (во генераторот за вгнездување) во зависност од пораката која што треба да ја пренесува. Главен фактор при менувањето на зборот е неговата прва буква т.е. во кодерот и декодерот се дефинира табела на мапирање (на пример *Табела 18*) според која бинарните вредности од 0000 до 1111 се мапираат во буквите од англиската азбука. Една бинарна вредност може да се мапира во повеќе букви.

Секој знак од скриената порака се претвора во бинарната репрезентација на неговиот *ASCII* код (прикажана како 8 бита) и потоа целата скриена порака се дели на групи од по четири бита. На пример:

Stop → 0101 0011 0111 0100 0110 1111 0111 0000

Со оглед на тоа дека група од четири бита се мапира во една буква (една буква одговара на еден клучен збор), а во примерот има 8 групи и со оглед на тоа дека секоја шега има еден клучен збор, потребно е генерирање на 8 шеги т.е. на 8 клучни зборови.

Табела 18: Мапирање за прикривање на четири бита во клучниот збор на шегата

Бинарна вредност	Прва буква на клучниот збор во шегата
0000	A
0001	B
0010	C
0011	D
0100	E
0101	F
0110	G
0111	H
1000	I
1001	J
1010	K
1011	L
1100	M
1101	N
1110	O
1111	P
0000	Q
0001	R
0010	S
0011	T
0100	U
0101	V
0110	W
0111	X
1000	Y
1001	Z

Користејќи го мапирањето на бинарните вредности во букви од англиската азбука, во кодерот и декодерот се дефинира алгоритамот според кој за секоја бинарна вредност се разгледуваат сите букви кои одговараат на бинарната вредност и се дефинира соодветен збор кој треба да почнува на една од буквите кој одговараат. Овој збор всушност претставува замена за оригиналниот клучен збор од шегата.

За групите од примерот даден погоре, се дефинира *Табела 19*.

Табела 19: Кодирана порака, користејќи ја првата буква од клучниот збор на шегата

Бинарна вредност	Буква	Клучен збор на шегата
0101	F или V	Vampire
0011	D или T	Teacher
0111	H или X	hamburger
0100	E или U	ugly
0110	G или W	www.square.com
1111	P	Public
0111	H или X	Hogwash
0000	A или Q	quarters

Врз основа на оваа табела, во генераторот за вгнездување се врши промена на клучните зборови со соодветните мапирани клучни зборови, што за примерот со скриената порака „Stop“ може да резултира во шегите:

- Where is Dracula's American office? The Vampire State Building.
- Teacher: When do astronauts eat? Pupil: At launch time!
- Can a hamburger marry a hot dog? Only if they have a very frank relationship!
- I'm not ugly. I could marry anyone I pleased! But that's the problem - you don't please anyone.
- Have you seen www.square.com? No, I haven't got around to it.
- What do you call 4 blondes laying on the beach? A: Public access.
- Why did the little pig hide the soap? He heard the farmer yell, "Hogwash!"
- Why is the moon like a dollar? It has four quarters.

Слика 8: Осум генерирани шеги, кои содржат 32 скриени битови со користење на клучните зборови на шегите

Секој знак од скриената порака креира две групи од по четири бита, што значи дека за секој знак од скриената порака е потребно генерирање на две шеги. Од овде се согледува малиот капацитет на методот и потребата за изнаоѓање на алтернативи на негово зголемување. Една алтернатива е користење на симболите кои луѓето денес најчесто го користат во нивната неформална комуникација, како што се дадени примерите во *Табела 20*.

Табела 20: Мапирање за прикривање на два битови со користење на симболи

Бинарна вредност	Симболи
00	☺ или ☹
01	:0) или :0(
10	:0)) или :0((
11	:-) или :-(

Симболите се дадени како мапирани вредности на соодветни бинарни вредности и истите се додаваат во рамките на шегите каде што истовремено се врши и претходно опишаната замена на клучните зборови.

Табела 21: Кодирана порака, користејќи ја првата буква од клучниот збор на шегата и користејќи симболи

Бинарна вредност	Буква	Клучен збор на шегата	Бинарна вредност	Симбол
0101	F или V	Vampire	00	☺
1101	D или T	Teacher	11	:-)
0100	H или X	hamburger	01	:0)
1011	E или U	ugly	10	:-)
0111	G или W	www.square.com	00	☺

На пример, табелата за дефинирање на мапирањето во кодерот и декодерот би се проширила како што е дадено во *Табела 21*.

Во овој случај, една шега може да биде преносител на повеќе од четири бита преку додавање на соодветни симболи на крајот на шегата, па скриената порака со 32 бита од примерот погоре, може да се групира на начин што истата ќе се вгнезди и во 6 генерирани шеги.

Stop → 0101 00 1101 11 0100 01 1011 11 0111 00 00

- Where is Dracula's American office? The Vampire State Building. ☺
- Teacher: When do astronauts eat? Pupil: At launch time! :-)
- Can a hamburger marry a hot dog? Only if they have a very frank relationship! :0)
- I'm not ugly. I could marry anyone I pleased! But that's the problem - you don't please anyone. :-)
- Have you seen www.square.com? No, I haven't got around to it. ☺ ☺

Слика 9: Шест генерирани шеги, кои содржат 32 скриени бита со користење на клучните зборови на шегите и со користење на симболи

Дополнително, една листа на шеги која е веќе преносител на скриена порака, може дополнително да се користи како „оригинална“ порака за вгнездување на друга скриена порака. На страната на кодерот и декодерот потребно е индексирање на нови клучни зборови и примена на истиот алгоритам како претходно.

На пример пораката:

war → 0111 0111 0110 0001 0111 0010

Вгнездена во пораката од *Слика 8* користејќи ја *Табела 18* и формирајќи мапа на соодветни клучни зборови, може да резултира во пораката од *Слика 10*.

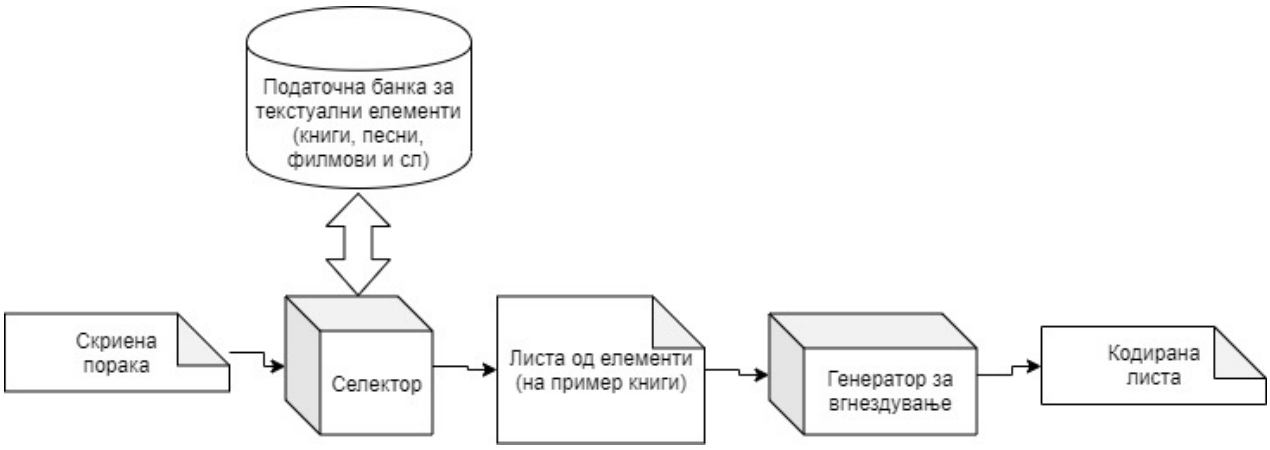
- Where is Dracula's American ~~office~~ house? The Vampire State Building.
- Teacher: When do astronauts eat? Pupil: Hmm At hunger time!
- ~~Can~~ Would a hamburger marry a hot dog? Only if they ~~have~~ retain a very frank relationship!
- I'm not ~~ugly~~ homely. I ~~could~~ can marry anyone I pleased! But that's the problem - you don't please anyone.

Слика 10: Четири генерирани шеги, кои содржат 24 скриени бита со менување на клучните зборови на шегите

2.4..2 Метод базиран на генератор на листи

Со методот [39] се врши генерирање на листи од елементи, кои се всушност преносители на скриени битови. Може да станува збор за генерирање на листа на книги, листа на песни, листа на филмови и слично. Овој начин на пренос може

да најде примена кога комуникацијата помеѓу две страни се одвива преку користење на онлајн платформи, кои навидум за крајните корисници се користат како промовирање на плејлисти со филмови/песни, нивно продавање, издавање и слично. Само корисниците кои знаат за постоењето на скриената порака и методот со кој се генерира листата, можат соодветно да ја декодираат пораката.

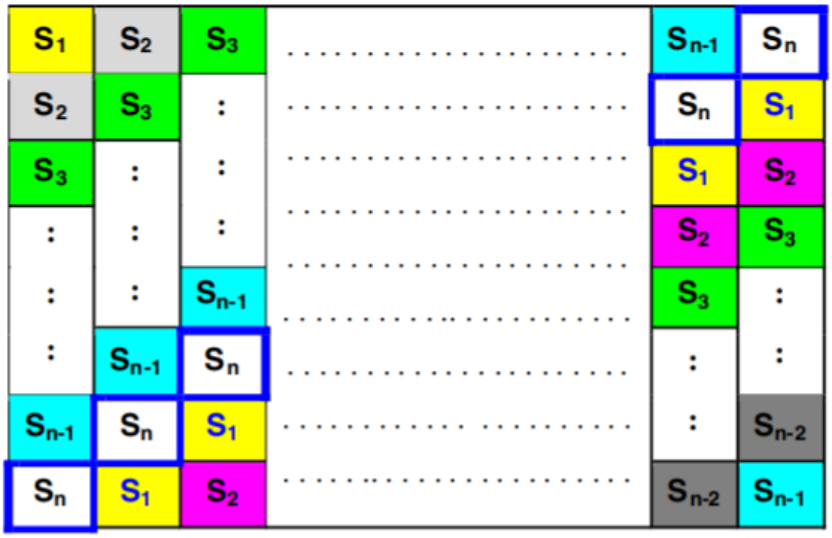


Слика 11: Илустрација на архитектурата за генерирање на листа за дадена скриена порака

Методот користи т.н. податочна банка со текстуални елементи која што всушност претставува база на податоци со текстуални имиња на книги, песни, филмови и слично. Важно е да се наспомене дека секој тип на елемент кој е составен дел од оваа база, потребно е да има по барем неколку елементи кои започнуваат на секоја буква од англиската азбука.

Модулот за селекција ја прима скриената порака во бинарен формат, врши селекција за тоа кој тип на листа ќе биде генериран и врз основа на одлуката зема соодветни елементи од податочната банка. Податоците понатаму се проследуваат во генераторот за вгнездување, каде што е сместен самиот алгоритам за генерирање на листите т.е. селектираните листи ги подредува на начин истите да вршат пренос на бинарни единици и бинарни нули.

Алгоритамот за вгнездување се базира на форматот на матрица т.н. *Латински квадрат* [40], каде што елементите се распоредуваат во *NxN* матрица на начин што секој елемент се содржи само еднаш во соодветниот ред и соодветната колона каде што е сместен.



Слика 12: NxN Латински квадрат – каде што секој ред / колона е уникатна пермутација од N елементи

Листата која што се генерира како резултат од алгоритамот содржи одреден број на елементи (книги / песни / филмови) во зависност од големината на бинарната порака која треба да се пренесе и во зависност од тоа како е дефиниран алгоритамот.

Првата буква на секој елемент ги дефинира битовите кој елементот ги носи, па во зависност од тоа колку битови носи еден елемент, се одредува колку елементи е потребно да се генерираат. Дополнително, се врши имплементирање на едно ниво на безбедност на алгоритамот (со цел да не биде лесно детектран од страна на крајниот корисник), така што првата буква на секој елемент ги дефинира битовите кој елементот ги носи, но мапирањето помеѓу буквите и битовите е променливо. Во текот на комуникацијата двете страни договараат начин на кој би разликувале различни итерации, и по секој премин од една итерација во друга, мапирањето се менува во согласност со претходно дефинираниот Латински квадрат. Користењето на оваа матрица овозможува еден вид на случајно генерирање на стеганографски код, но притоа сè уште се задржува одредена шема според која примателот на пораката ќе може успешно да ја декодира скриената порака.

Во продолжение е даден едноставен пример за тоа како функционира мапирањето, каде што е дадена соодветната матрица со кој се дефинира преносот на два бита, притоа користејќи четири букви.

Табела 22: Користењето на четири букви за пренос на два бита преку кружно менување на шемата во четири итерации

Прва буква	A	B	C	D
Прво користење	00	01	10	11
Второ користење	01	10	11	00
Трето користење	10	11	00	01
Четврто користење	11	00	01	10

Во случајот, доколку на пример е потребно да се пренесе буквата „X“ како скриена порака, за истата се наоѓа соодветниот бинарен *ASCII* формат:

X → 01011000 → 01 01 10 00

Според дефинираната *Табела 22*, потребно е понатаму пораката се дели на групи од по два бита што резултира со 4 групи т.е. ќе биде потребно генерирање на лист со четири елементи. Секој од тие елементи ќе треба да започнува на една од буквите A, B, C, D во зависност од тоа во која итерација е комуникацијата меѓу двете страни. За конкретниот пример, случајното генерирање е дадено во *Табела 24*.

Табела 23: Демонстрација на ефектот на случајно генерирање, со користење на Latin Square матрица

Во првата итерација елементите на генерираната листа треба да започнуваат со буквите	Во втората итерација елементите на генерираната листа треба да започнуваат со буквите	Во третата итерација елементите на генерираната листа треба да започнуваат со буквите
01 = B	01 = A	01 = D
01 = B	01 = A	01 = D
10 = C	10 = B	10 = A
00 = A	00 = D	00 = C

Во *Табела 24* е дадена целосната матрица која се користи во крајната верзија на овој метод, на начин што се искористуваат сите 26 букви од англиската азбука за пренос на четири бита по елемент од листата.

Табела 24: Користењето на сите букви од англиската азбука за пренос на четири бита преку кружно менување на шемата во дваесет и шест итерации

Бинарен код	00000	00001	00010	00011	00100	00101	00110	00111	01000	01001	01010	01011	01100	01101	01110	01111	10000	10001	10010	10011	10100	10101	10110	10111	11000	11001
	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
2	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A
3	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B
4	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C
5	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D
6	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E
7	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F
8	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G
9	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H
10	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I
11	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J
12	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K
13	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L
14	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M
15	O	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N
16	P	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
17	Q	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
18	R	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
19	S	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
20	T	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S

21	U	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
22	V	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
23	W	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
24	X	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
25	Y	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
26	Z	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y

Во *Табела 25* е даден конкретен пример за генерирање на листа од песни во првата итерација од комуникацијата за скриената порака:

get him → 0110 0111 0110 0101 0111 0100 0010 0000 0110 1000 0110 1001 0110 1101

Во *Табела 26* е даден конкретен пример за генерирање на листа од книги во втората итерација од комуникацијата за скриената порака:

Stop → 0101 0011 0111 0100 0110 1111 0111 0000

Табела 25: Генерирање на листа со песни која врши пренос на скриената порака „get him“ во првата итерација од комуникацијата

Бинарен код	Прва буква на редот, користејќи ја првата итерација	Листа на песни
0110	G или W	*Wicked Game-1989 Chris Isaak lyrics*
0111	H или X	*How Do I Live-1997 LeAnn Rimes lyrics*
0110	G или W	*Wonderful Tonight-1978 Eric Clapton lyrics*
0101	F или V	*Faithfully-1983 Journey lyrics*
0111	H или X	*Hero-2001 Enrique Iglesias lyrics*
0100	E или U	*Endless Love-1981 Diana Ross & Lionel Richie lyrics*
0010	C или S	*Careless Whisper-1984 Wham! Lyrics*
0000	A или Q	*And I Love Her-1964 The Beatles lyrics*
0110	G или W	*Wild Thing-1966 The Troggs lyrics*
1000	I или Y	*Your Song-1971 Elton John lyrics*
0110	G или W	*Girl-1965 The Beatles lyrics*
1001	J или Z	*Just Fine" is a song by Mary J. Blige*
0110	G или W	*I Want You To Want Me-1979 Cheap Trick lyrics*
1101	N	*Nobody Wants To Be Lonely-2000 Ricky Martin lyrics*

Табела 26: Генерирање на листа од книги која врши пренос на скриената порака „Stop“ во втората итерација од комуникацијата

Бинарен код	Прва буква на редот, користејќи ја втората итерација	Листа на книги
0101	G или W	*Warrior Heir, (2006). Axelrad, Catherine*
0011	E или U	*Ever (2008). Fitzgerald, M.*
0111	I или Y	*Year of Fog, (2008). Scott Sigler*
0100	F или V	*Vengeful Virgin, (1958). Benjamin, Ross*
0110	H или X	*Hunting Wind, (2002). Smith, Melissa*
1111	Q	*Q is for Quarry (2002). Sue Grafton*
0111	I или Y	*Inventing the Abbotts (1987). Joss, Morag*
0000	B или R	*Blood Is the Sky (2004). Steve Hamilton*

Под премин од една итерација во друга може дури да се дефинира дури и секој нареден знак или секој нареден збор од скриената порака, така што се методот се обезбедува до степен што доколку некоја трета страна забележи дека методот се користи, доколку не го знае преминот од една итерација во друга, сепак нема да може да ја декодира скриената порака.

2.5 Постоечки методи за стеганализа

Стеганализата (*steganalysis*) преставува процес за детекција на тоа дали еден документ е преносител на скриена порака. Главна задача на повеќето методи на стеганализа е да пронајдат разлики во дистрибуцијата на одредени статистички особини помеѓу легитимните носачи и стего-носачите. На пример, ако носачот е текстуален документ, постоење на одредена шема на повторување на грешки, несекојдневни промени на одредени ентитети (знаци, зборови, линии, реченици, параграфи и слично) би можеле да преставуваат индикации дека нешто не е во ред со документот. Според примената, постојат **специфични методи** кои адресираат одредена стеганографска метода, и **универзални методи**, кои се обидуваат да се справат со повеќето или со сите стеганографски техники. Додека првите постигнуваат голема точност на детекција во пракса, вторите се поатрактивни, бидејќи не зависат од самиот стеганографски метод, и дури можат да се применат на непознати или идни стеганографски методи.

Крајниот резултат од извршената стеганализа може да има различни формати, но крајната цел е секогаш поврзана со постоењето на скриена порака. Како на пример, резултатот може да биде даден со директен нумерични формат т.е. процентуална проценка за тоа колку се шансите за тоа да постои скриена порака. Друг формат е прикажување на одредени статистики поврзани со документот кој се анализира (броење на одредени невообичаени својства на ентитетите, броење на одредени грешки и слично) и врз основа на тие статистики е оставено крајниот корисник сам да ги анализира и самостојно да донесе заклучок за тоа дали мисли дека постои скриена порака или не.

Анализирањето на документот обично се врши земајќи предвид одреден тип на стеганографски методи, при што претежно се бројат / проверуваат / проценуваат оние својства на документот кои генерално се менуваат со множеството на стеганографските методи кои притоа се земени предвид. Од овде произлегува постоењето на различни методи за стеганализа, при што секој метод врши детекција на тоа дали постои скриена порака која што е вгнездена со користење на одредено множество од стеганографски методи. Една од целите на секој метод за стеганализа е да може да детектира што е можно поголем број на стеганографски методи. За секој стеганографски метод може да се развијат соодветни методи на стеганализа на начин што дел од тие методи може да даваат директна процентуална проценка, а дел од нив може да прикажуваат одредени статистики.

Развивањето на методите на стеганализа е дел од еден многу широк спектар на можности, со оглед на тоа дека не постои одредена шема и одреден начин на кој што се врши детекцијата. Алгоритмите при користење на стаганализата зависат од алгортимите користени при стеганографијата и при развивањето на истите не се следат некои одредени правила т.е. начинот на пристап за добивање на крајниот резултат и начинот на прикажување на него може да е различен за секоја индивидуа која што ја врши стеганализата.

Доколку крајниот резултат од една стеганализа укажува на сигурност кон тоа дека не постои скриена порака во документот, таа сигурност може да се интерпретира како сигурност дека не постои вгнездена скриена порака со користење на стеганографските методи кои ги користат својствата коишто се анализираат во методот на стеганализа. Од друга страна пак, таа сигурност не

може да биде показател кон тоа дека во документот не постои никаква скриена порака, со оглед на тоа дека постојат најразлични стеганографски методи па речиси е невозможно да се изврши една глобална проценка која што би ги опфатила сите стеганографски можности.

Во зависност од типот на ентитетот кој се користи за пренос на скриените пораки, постојат повеќе видови на стеганализа. За секој тип на стеганографија, постои соодветен тип на стеганализа, како на пример: стеганализа во дигиталните медиуми (слики, аудио документи, видео документи и слично), стеганализа во системот со датотеки (filesystem), мрежна стеганализа, текстуална стеганализа итн.

Текстуалната стеганализа е еден од најкористените облици на стеганализа (со оглед на тоа дека соодветствува на текстуална стеганографија) и таа се обидува да идентификува дали одредена текстуална порака/текстуален документ содржи скриена порака. Развојот на областа на текстуалната стеганализа е од голема важност, бидејќи токму текстуалните документи претставуваат голема закана како потенцијални преносители за комуникација меѓу сајбер-криминалци и меѓу терористи. Друга корисна примена на стеганализата е проверка на авторски права („*copyright*“), верификација на содржини, следење на документи и слично.

Текстуалната стеганализа се класифицира во три категории: лингвистички методи за стеганализа, методи за стеганализа базирани на невидливи знаци и методи за стеганализа базирани на формати.

Заедничко за методите од сите типови е тоа што повеќето од нив користат статистички пресметки и *SVM* модели за класификација. *SVM (support-vector machines)* се надгледувани модели за учење, кои се поврзани со алгоритми за учење и како краен резултат вршат класификација на податоци во една или од повеќе категории. Овие модели користат претходно одредени множества со примери од текст, каде што секој пример е веќе означен во соодветна категорија во која што припаѓа. Алгоритамот ги испитува својствата на примерите на начин што по извршено тренирање со користење на поголем број примери, алгоритамот потоа самиот станува способен да означува категории на нови примери кои нема да бидат означени.

Постојат лингвистички методи за стеганализа кои вршат детекција на поголемо множество од стеганографски лингвистички методи, при што со користење на тренирани лингвистички модели и *SVM* модели за класификација [49] се детектираат лексичките линквистички методи, или пак со вгнездување на зборови се детектираат семантички нарушувања предизвикани од користење на синоними [50] и слично.

Поголемиот број на методи за стеганализа кои се наменети за детекција на стеганографски методи базирани на случајно и статистичко генерирање, користат конволуциски невронски мрежи [51] [52], би-дирекционални невронски мрежи со повторување [53], конволуциски поместувачки прозори (*TS-CSW*) [54] или семантичка поврзаност помеѓу зборовите заедно со *softmax* класификатор [55] со цел земање на површинската семантика на текстот или површинската поврзаност помеѓу зборовите, за изнаоѓање на разликите во семантичкиот простор пред и по вгнездувањето на скриената порака.

2.6 Лингвистички методи за стеганализа

Лингвистичките методи за стеганализа вршат анализа за потенцијалниот пренос на скриени пораки, со користење на стеганографските лингвистички методи. Со оглед на комплексноста на лингвистиката, тие вообичаено се наменети само за одреден јазик (најчесто англискиот) и често користат предефинирани множества / тренирани модели, кои се креирани врз база на текстови со точна граматика (книги, песни и слично).

2.6..1 Метод базиран на статичките карактеристики на врските помеѓу зборовите

Овој методот е наменет за одредување на веројатноста за тоа дали документот содржи скриена порака, вгнездена со користење на некои од стеганографските лингвистички методи кои што ги менуваат зборовите од оригиналниот документ, со што се менува значењето и на самата содржина во документот. На пример, од набројаните методи во претходните глави, овој метод за стеганализа може да изврши детекцијата доколку е користен стеганографскиот метод за следење на промените на зборовите или пак методот за мапирање на зборовите.

Методот се базира на врската помеѓу зборовите во рамките на една реченица, која што е пресметана според соодветен алгоритам, притоа користејќи голем број на постоечки книги, романи и текстови, како множество од реченици кое претставува т.н. множество за тренирање. Идејата е алгоритамот да научи да детектира кога реченицата има природно и комплетно значење. На пример, за реченица која што започнува со зборовите:

She is a . . .

Алгоритамот би очекувал таа да завршува со зборови кои веќе имаат воспоставено одредени врски со зборовите на кои започнува реченицата, при користење на множеството за тренирање. На пример:

She is a woman teacher.

She is a beautiful actress.

She is a mother.

Појавата на зборови кои немаат воспоставено соодветни врски со првите зборови од реченицата, ќе бидат детектирани како неочекувана појава. На пример:

She is a man.

She is a good actor.

She is a father.

По користењето на алгоритамот за одредување на врските, за конкретните дадени примери, алгоритамот „научил“ дека зборот „*she*“ или воспоставено силна врска со зборовите „*woman*“, „*actress*“ и „*mother*“, па токму затоа тој очекува дека овие зборови е многу веројатно да се појават во продолжението на реченицата. Од друга страна пак, врската на зборот „*she*“ со зборовите „*man*“, „*actor*“ и „*father*“ е многу слаба или пак не е ни воспоставена. Интензитетот на овие врски е изразен преку нумерички вредности, на начин што колку е посилна врската помеѓу два збора, таа е изразена со поголема вредност.

За подобар опис на методот, авторите на [41] воведуваат неколку термини:

- *MI (Mutual Information)* – врската дефинирана помеѓу два конкретни зборови:

$$MI(x,y) = log_2 \frac{P(x,y)}{P(x)\,P(y)}$$

каде што *P*(*x,y*), *P*(*x*) и *P*(*y*) се веројатностите за појава на „xy“, „x“ и „y“ во дадениот текст, соодветно;

- *N-WWP (N-Window Word Pair)* – пар на зборови во рамките на една реченца, со растојание помало од N (каде што N>1). На пример, за реченицата: „*word1 word2 word3 … word10*“ доколку N=4, секој пар од два збора помеѓу зборовите „word1 – word4, word2 – word5 … word7 – word10“, претставува *4-WWP* пар во реченицата;
- *N-WC (N-Window Collocation)* – претставува *N-WWP* со голема фреквенција;
- *N-WMI (N-Window Mutual Information)* – претставува врската дефинирана помеѓу два збора во рамките на *N-WWP*. Дополнително, еден *N-WWP* претставува *N-WC* само доколку вредноста на *N-WMI* е поголема од вредноста дефинирана како:

$$MI_N(x,y) = log_2 \frac{CC_{xy}}{C_x C_y}$$

каде што *C*, C_{xy}, C_x и C_y го претставуваат бројот на појавувања на *N-WWP*, (x,y), (x,?) и (?,y), соодветно. На пример, за реченицата:

We were testing, but testing failed.

За евалуација на врската за парот (*were,testing*) потребно е да се разгледат сите *4-WWP* од реченицата: (*we,were*), (*we,testing*), (*we,but*), (*were,testing*), (*were,but*), (*were,testing*), (*testing,but*), (*testing,testing*), (*testing,failed*), (*but,testing*), (*but,failed*), (*testing,failed*). Со анализа на паровите, се добиваат вредностите:

C=12 – бројот на парови;

$C_{were,testing}=2$ – бројот на појавувања на конкретниот пар;

$C_{were,?}=3$ – бројот на парови каде што *were* е прв збор во парот;

$C_{?,testing}=5$ – бројот на парови каде што *testing* е втор збор во парот;

со користење на формулата за *N-WMI* ќе се добие јачината на врската прикажана во нумеричка вредност и примерот всушност го прикажува процесот на пресметка на врските помеѓу зборовите.

Во методот се користи генерален речник од *M* зборови и се дефинира матрица *MxM* каде што секој пресек на колона / редица претставува соодветен пар од два збора. Потоа се користи значителен број на множество за тренирање (голем број на текстови) каде што се пресметуваат врските помеѓу зборовите на начинот прикажан погоре. Целта е на крајот од процесот да се добијат соодветните *N-WMI* вредности за секој пар на зборови (за секој пресек) од матрицата.

Кога процесот за стеганализа за одреден документ е започнат, врз основа на вредностите од матрицата *MxM* алгоритамот ги наоѓа нумеричките вредности за врските помеѓу зборовите во документот и одредува дали истите се слаби или силни. Овие вредности претставуваат основа за пресметки на два дополнителни параметри преку кои потоа се одредува веројатноста за постоењето на скриена порака.

Од текстот на документот кој се анализира, следејќи го примерот погоре, се формира матрица S_{MxM}. Од друга страна пак, нека од матрицата со генералниот речник и со добиени вредности врз основа на тренирањето се извлечат соодветните зборови во редиците / колоните и нека матрицата со вредностите добиени од тренингот се означи како T_{MxM}. Еден од параметрите за детекција се дефинира како:

- *N-WVMI (N-Window Variance of Mutual Information)*

$$V = \frac{1}{MxM}\sum_{i=0}^{M}\sum_{j=0}^{M}(S_{ij} - T_{ij})^2$$

каде што со *S* и *T* дадени матриците споменати погоре. Другиот параметар е дефиниран како:

- *PAD (Partial Average Distance)*

$$D_{\alpha,K} = \frac{1}{K}\sum_{i=0}^{M}\sum_{j=0}^{M}|S_{ij} - T_{ij}|[|S_{ij} - T_{ij}| > \alpha]\lambda_K\,(i,j)$$

каде што α е предефинирана вредност и го преставува растојанието помеѓу две *N-WMI* вредности и *K* е предефинирана вредност на зададено ограничување

дека само првите K најголеми вредности од матрицата S_{MxM} ќе бидат пресметани.

Авторите на трудот [41] дошле до овие два параметра преку различни статистички операции и истите ги користат како влезни параметри во *SVM* модел кој што врши крајна директна класификација за тоа дали документот кој се анализира содржи скриена порака или не.

Горе дадените пресметки претставуваат површинско разгледување на крајните резултати од алгоритамот и процесот, додека процесот да се дојде до самите формули е покомплексен и вклучува подетално познавање од областа на статистиката и машинското учење.

2.6..2 Метод базиран на мета-својства и механизам на имунизација

Овој метод се базира на некои статистички својства на лингвистиката на англискиот јазик. Авторите на [42] дефинираат 57 основни статистички својства и истите ги нарекуваат „мета-својства“. Преку различни анализи и експерименти, донесени се статистички заклучоци за тоа какви вредности имаат својствата применети во нормален документ споредено со тоа какви вредности имаат кога се применети во документ во кој е вгнездена скриена порака.

Како на пример, дел од овие својства се:

- просечната должина на зборовите во документот – документите со скриен текст вообичаено имаат подолга просечна должина на зборовите од нормалните документи;
- стапката на празен простор – документите со скриен текст вообичаено имаат помалку празен простор од нормалните документи;
- зборови кои имаат висока фреквенција (*AFW*) и зборови кои имаат ниска фреквенција (*UFW*) – документите со скриен текст вообичаено имаат помалку зборови со ниска фреквенција од нормалните документи;
- процент на буквите – документите со скриен текст вообичаено имаат помалку промени на буквите од нормалните документи;
- дистрибуција на првите букви – кај документите со скриен текст, дистрибуцијата на првите букви е порамномерна од онаа кај нормалните документи; итн.

Со цел донесување на соодветни заклучоци за „мета-својствата“ на документите, користено е големо множество од документи. По извршено броење на својствата, дојдено е до заклучок дека документите кои имаат вгнездени скриени пораки, имаат помали стапки на разгледаните „мета-својствата“ од нормалните документи (како што може да се забележи и од опишаните својства погоре). За таа цел се воведува терминот на имун механизам за клонирање, каде што секој текст може да се претстави како:

$$v = (v_1, v_2, \dots, v_{57})$$

каде што v_i е вредност која што го претставува *i*-тото „мета-својство“. Терминот на имун механизам е претставен во [43] и истиот има улога да генерира множество од детектори, кои според одредени својства вршат класификација на тоа дали еден текст би можел да содржи скриена порака или не. Колку една скриена порака е поголема, толку нејзините „мета-својства“ имаат помали вредности. Текстот кој има вгнездена скриена порака, според „мета-својствата“ се дели на две категории:

- *SST (Success-Stego-Text)* – текст кој содржи вгнездена порака и на кој „мета-својствата“ одговараат на вредности кои се специфични за текст во кој има вгнездена порака;
- *FST (Failure-Stego-Text)* – текст кој содржи вгнездена порака и на кој „мета-својствата“ одговараат на вредности кои се специфични за текст во кој нема вгнездена порака.

Во рамките на методот прво се генерираат детектори на *SST* текст и со имун механизам тие детектори се тренираат. Потоа истото тоа се врши и за *FST* текст, прво се генерираат детектори, а потоа тие се тренираат со користење текстови со скриени пораки и со нормални текстови.

Откако детекторите се генерирани и креирани, како влезен параметар во методот се пушта документот кој што треба да се анализира, а истренираните детектори (врз основа на „мета-својствата“ на текстот) го класифицираат во една од категориите, дали содржи скриена порака или не.

Горедадените пресметки претставуваат површинско разгледување на крајните резултати од алгоритамот и процесот додека да се дојде до самите формули е покомплексен и вклучува подетално познавање од областа на статистиката и на машинското учење.

2.7 Методи за стеганализа базирани на невидливи знаци

2.7..1 Метод базиран на семантички простор на конволуциска невронска мрежа

Овој метод [44] е базиран на семантичка анализа која што користи конволуциска невронска мрежа (*CNN*) за извлекување на семантичките својства на текстот и изнаоѓање на разликите во дистрибуцијата на семантичкиот простор пред и по вгнездувањето на скриената порака.

Секоја реченица која што е влезен параметар во анализата, се претставува со матрица каде што *i*-тиот ред го означува *i*-тиот збор и секој збор е претставен како вектор. Ако се разгледува $X_{i:j}$ како матрица од вектори на зборови од *i*-тиот збор до *j*-тиот збор, тогаш може да се извлече својството:

$$c_i = f(\mathrm{W} * X_{i:j} + b_i)$$

каде што *f* е нелинеарна функција и b_i е степенот на шум. Низата од сите својства за матрицата, ја дефинира мапата на својства, која што подоцна може да резултира со „вектор на својства“.

Има два модели на овој метод:

- *TS-CNN (Single)* – каде што сите реченици од текстот се споени во една линија и овој пристап привлекува повеќе внимание на семантиката помеѓу реченициите како една целина;
- *TS-CNN (Multi)* – каде што сите реченици од текстот се третирани како посебни линии и овој пристап привлекува повеќе внимание на семантиката во самите реченици, за секоја реченица посебно;

Алгоритамот резултира со формулата:

$$O = S(W_F \, \mathrm{z} + b_f)$$

каде што *S* е сигмоидна функција, *z* е всушност “векторот на својства”, W_F и b_f се истренираната матрица и истренираниот шум, соодветно. На крајот на алгоритамот се дефинира одредена вредност – *t*, на начин што крајната детекција резултира со:

$$\text{Резултат} = \begin{cases} \text{Текст со вгнездена скриена порака}, & O \geq t \\ \text{Текст што не содржи скриена порака}, & O < t \end{cases}$$

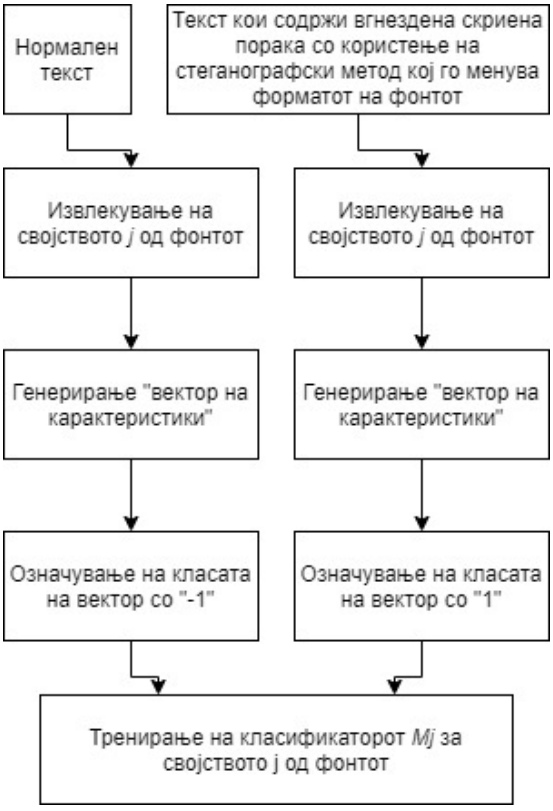
Горе дадените пресметки претставуваат површинско разгледување на крајните резултати од алгоритамот и процесот додека да се дојде до самите

формули е покомплексен и вклучува подетално познавање од областа на статистиката и на машинското учење.

2.8 Методи за стеганализа базирани на формати

2.8..1 Метод базиран на формати на фонт

Методот [45] користи *SVM* модел како класификатор за детекција на постоењето на скриена порака, вгнездена со методи кои ги менуваат својствата на фонтот. Се врши анализа на својствата на фонтот на секој знак и за секое својство се креира „вектор на карактеристики“. Како множество за тренирање на моделот се користат нормални текстови – текстови без вгнездени скриени пораки, како и истите тие текстови по вгнездувањето на скриена порака со соодветен стеганографски метод (кој го менува форматот на фонтот). Два “вектори на карактеристики” (кои одговараат на истото својство на истиот знак) се означуваат со различни класи и двата резултати се праќаат како влезен параметар во класификаторот со цел истиот да се тренира.



Слика 13: Процес на тренирање на класификатор M_j

Секој знак има повеќе својства на фонтот. Ако секое нормално (непроменето) својство се мапира како константна нумеричка вредност $p(t_i)$, каде што t_i претставува знак во текстот и ако двете категории од класификаторот се означени со позитивна/негативна нумеричка вредност $\gamma(t_i)$, тогаш за секој знак може да се пресмета секвенцата $\{P(t_i)\}$, каде што:

$$P(t_i) = p(t_i) + \gamma(t_i) + \delta(t_i)$$

вредноста $\delta(t_i)$ претставува сигнал на шум – таа варира во зависност од вредноста на својството на фонтот и е предизвикана од извршените промени врз оригиналните својства на фонтот. Растојанието помеѓу својствата два соседни знаци t_i и t_{i+1} е дефинирано како:

$$D_i = |P(t_i) - P(t_{i+1})| = |p(t_i) - p(t_{i+1}) + \gamma(t_i) - \gamma(t_{i+1}) + \delta(t_i) - \delta(t_{i+1})|$$

од каде што фреквенцијата на растојанието помеѓу својствата е бројот на пати колку што дадена вредност за растојанието се појавува во $\{D_i\}$ и истото е изразено како *S*:

$$S_k = \sum_{i=0}^{n-2} (D_i = k)$$

секвенцата $\{D_i\}$ може да се подели во помали области во зависност од λ (граница на растојанието помеѓу својствата, која што ги дели едни од други) и колку што λ има поголема вредност, поголеми се и овие помали области. Фреквенција на вкупното растојание во секоја област може да се пресмета како „вектор на карактеристики“ од *SVM*.

Проценатa на должината на скриената порака е:

$$L = C \; x \; \beta \; x \; \frac{1}{\alpha} \; (bit)$$

каде што C е бројот на ненултите елементи во $\{D_i\}$, β е бројот на битови кои се вгнездени во секој знак, а α е параметар кој методот го генерира и го претставува односот на вкупниот број на битови во една подобласт кои не се еднакви со битовите од соседната подобласт, со вкупниот број на битови во даден текст.

Горе дадените пресметки претставуваат површинско разгледување на крајните резултати од алгоритамот и процесот додека да се дојде до самите

формули е покомплексен и вклучува подетално познавање од областа на статистиката и на машинското учење.

2.8..2 Метод базиран на поместување на зборови

Метод [46] извршува проценка за тоа дали во документот постои скриена порака со користење на стеганографските методи кои вршат поместување на зборовите. Доколку множеството од сите зборови во документот се прикаже како:

$$W = \{W_n\}_{n=1}^{w}$$

и доколку празниот простор помеѓу зборовите W_i и W_{i+1} се престави како S_i, тогаш множеството со празни места може да се прикаже како:

$$S = \{S\}_{n=1}^{w-l-e}$$

каде што l е бројот на линии во документот (бидејќи нема празен простор помеѓу две соседни линии, тие празни места не се вклучени) и e е бројот на реченици (под претпоставка дека нема празни места кога завршуваат реченициите).

Се воведува концептот на соседни растојанија, кој е дефиниран како разликата помеѓу должините на две соседни празни места:

$$D_i = S_{i+1} - S_i$$

доколку S_i и S_{i+1} се во иста линија. Множеството од соседни растојанија е преставено како:

$$D = \{D_n\}_{n=1}^{w-2l-e}$$

дополнително, хистограмот на соседни разлики е дефиниран како:

$$H(i) = \sum_{t=0}^{w-2l-e} \delta(D_t, i), \;\; i = \ldots, -0.2, -0.1, 0, 0.1, 0.2, \ldots$$

каде што $\delta(x, y)$ е функција која враќа само две вредности:

$$\delta(x, y) = \begin{cases} 1, & y - 0.05 \leq x \leq y + 0.05 \\ 0, & \text{во спротивно} \end{cases}$$

Под претпоставка дека S_i има постојана вредност за t, скриената порака е независна и рамномерно распределена и под претпоставка дека битовите од скриената порака се вгнездени *R* битови по ентитет-преносител, тогаш вкупниот

број на непроменети празни места (при процесот на вгнездување) може да се пресмета како:

$$\widehat{H}(\pm\mu) = \frac{R}{2}\|D\|, \qquad \widehat{H}(0) = (1-R)\|D\|$$

каде што μ е стеганографски параметар на начин што вредноста на соседните разлики се движи во рангот од *{- μ, 0, + μ}*.

Разгледаниот хистограм H може да се спореди со $\widehat{H}$, со користење на *chi-square* статистиката [47]:

$$x^2 = \sum_{t=-\mu,0,\mu} \frac{(\widehat{H}(t) - H(t))^2}{\widehat{H}(t)}$$

Доколку текстуалниот документ не бил променет (хистограмот H не е поврзан со $\widehat{H}$) тогаш следниот услов ќе биде исполнет:

$$x^2 > {x_\alpha}^2$$

Каде што параметарот α може да биде земан како вредност 0.05 и параметарот μ може да биде земан како вредност *3pt*.

Горе дадените пресметки претставуваат површинско разгледување на крајните резултати од алгоритамот и процесот додека да се дојде до самите формули е покомплексен и вклучува подетално познавање од областа на статистиката.

3 Резултати

По извршената анализа на постојните методи за стеганографија и стеганализа, во продолжение се разгледани нови методи кои се креирани во склоп на оваа магистерска теза и при тоа, за истите има соодветна имплементација која подетално ќе биде разгледана во делот за дискусија.

3.1 Предложени методи за стеганографија

Предложени се четири нови стеганографски методи базирани на формати, за криење на пораки во *Microsoft Word* документи. Предложените методи вршат манипулации со различни својства на текстот кои не се видливи за човековото око и тие се класифицирани во една поткатегорија, наречена „методи базирани на својствата на текстот“ [48].

3.1..1 Скалирање на знаците

При користење на софтверот *Microsoft Word* , знаците на текстот имаат својство за скалирање кое секогаш има пред-дефинирана вредност 100%. Оваа вредност крајниот корисник може да ја менува, но генерално тоа е многу малку веројатно и секој новокреиран текстуален документ во *Microsoft Word*, резултира со вредност за скалирање од 100% за текстот кои почнува да се пишува. Менувањето на оваа вредност ја променува големината на знаците, на пример зголемувањето/намалувањето на скалирањето на буквите ги прави поголеми / помали, додека пак зголемувањето/намалувањето на скалирањето на знакот за празно место го зголемува/намалува празниот простор помеѓу знаците од левата и од десната страна на тој знак.

Вршењето на поголеми промени на вредноста за скалирањето е препознатливо за човековото око, но вршењето на мали промени може да помине незабележително. На пример, доколку некои од знаците имаа вредност за скалирање од 99%, а други имаат вредност од 100%, знаците и да се наоѓаат еден до друг, човековото око тоа нема да може да ги забележи тие промени со оглед на тоа што разликата од 1% е многу мала.

Предложениот метод претпоставува користење на две различни вредности за скалирање, блиски една до друга за преставување на бинарни единица и нула. На пример, ако пред-дефинираната вредност за скалирање на знаците е

100%, еден начин е на пример, 99% да се користи за бинарна единица, а 100% да се користи за бинарна нула. Или ако сакаме во текстот не сите знаци да носат скриена порака, тогаш втората вредност наместо 100% може да се земе на пример, да биде 101%. Во тој случај, знаците со 100% нема да пренесуваат скриени битови.

Методот е тестиран, при што е забележано дека постоењето на овие три скалирања врз знаци кои се еден до друг, не е забележливо од крајниот корисник и методот успешно може да пренесува скриени пораки, без крајниот корисник да забележи дека нешто е менувано во документот.

Капацитетот на овој метод зависи од бројот на знаци во документот, со тоа што секој знак (буква, празно место, интерпункциски знак) е потенцијален преносител на еден бит.

Предложени се и други алтернативи на методот, но истите не се вклучени во практичниот дел со имплементацијата. Една модификација е да се земат четири вредности на скалирање кои се блиску една до друга, и тогаш секој знак може да пренесува два бита. На пример:

- знаците кои имаат вредност за скалирање од 97% претставуваат ентитети за пренос на 00;
- знаците кои имаат вредност за скалирање од 98% претставуваат ентитети за пренос на 01;
- знаците кои имаат вредност за скалирање од 99% претставуваат ентитети за пренос на 10;
- знаците кои имаат вредност за скалирање од 100% претставуваат ентитети за пренос на 11.

Со експерименти е потврдено дека и оваа алтернатива, исто така може да помине незабележително од страна на крајниот корисник, но со самото користење на вредности кои се „подалеку“ од вредноста 100%, методот се доведува во поголем ризик да биде забележан.

Друг фактор кои може да влијае на забележливоста на методот е фактот дека знаци со различни вредности на скалирање се сместени едни до други. За намалување на ризикот постои друга алтернатива, каде што наместо менувањето на скалирањето на секој знак, се менува скалирањето на секој збор. Со тоа што зборовите се одвоени со празно место, придонесува различните

скалирања да не се сместени директно едни до други. Доколку скалирањето се применува врз зборовите, капацитетот на методот значително се намалува, со тоа што секој збор претставува потенцијален преносител на еден бит. И кај зборовите важат опциите како кај знаците т.е. доколку се користат две нови вредности за скалирање, секој збор е преносител на еден бит, а доколку се користат четири нови вредности за скалирање, секој збор е преносител на два бита.

3.1..2 Подвлекување на знаците

Едни од основните својства на текстот во *Microsoft Word* документите се задебелувањето на знаците (**bold**), закосувањето на знаците (*italic*) и подвлекувањето на знаците (underline).

Предложениот метод директно го искористува својството за подвлекување на знаците, на начин што крајниот корисник тоа нема да го забележи. Знаците се подвлекуваат и на самата линија за подвлекување се менува стилот и бојата, што резултира со повеќе различни комбинации кои потоа може да се интерпретираат како преносители на скриени битови.

При подвлекувањето на знаците, постојат 16 различни стила за подвлекување кои корисникот може да ги избере, со што стилот на линијата за подвлекување овозможува пренос на четири бита. Преносот се врши со дефинирање на мапа т.е. мапирање на секој стил со соодветна вредност од 0000 – 1111, па при процесот на декодирање се користи таа мапа и при читањето на стилот на линијата за подвлекување на секој знак, се добива соодветната вредност од четири бита.

Со цел линијата за подвлекување да помине незабележително, методот претпоставува дека од аспект на бојата на позадината на документот и бојата на позадината на самите знаци, се користи најчестиот формат на *Microsoft Word* документи каде што позадината е бела т.е. без боја. На линија за подвлекување, исто така се задава бела боја, на начин што таа се аплицира само на линијата за подвлекување, но не и на самиот знак, со што линијата за подвлекување се синхронизира со позадината и не е видлива за човековото око. Манипулацијата со бојата на линијата може да има и две намени, па освен тоа што таа се користи за прикривање на подвлекувањето, таа може да биде и преносител на скриени

битови. Белата боја претставена преку *RGB* параметрите ја има вредноста (255, 255, 255), но преку извршени експерименти е дефинирано дека најблиските нијанси до оваа вредност се толку блиски и светли што човековото око не може да ги препознае. За таа цел се дефинираат најблиските 16 вредности на бела боја (*R,G,B* >= 253) и со тоа изборот на боја на линијата за подвлекување исто така станува преносител на четири бита.

Со тоа што изборот на стил овозможува пренос на четири бита и изборот на боја овозможува пренос на четири бита, секој знак станува потенцијален преносител на 8 битови т.е. методот има многу голем капацитет на пренос. Друга предложена алтернатива на методот е изборот на една од најблиските 216 вредности на белата боја (*R,G,B* >= 249) со оглед на тоа дека експериментите покажуваат дека нијанси на белата боја и во тој случај се прифатливи и непрепознатливи за човековото око. Од тие 216 комбинации, би можеле да се искористат најблиските 128 комбинации, што би овозможило пренос на 7 бита преку бојата на линијата за подвлекување и 11 бита (7+4) по знак. И покрај тоа што оваа алтернатива овозможува поголем капацитет за пренос, фактот дека се избираат нијанси на бела боја кои се „подалеку“ од основната нијанса на бела боја, доведува до поголема изложеност на алгоритамот да може да биде препознаен од страна на крајниот корисник.

Знаците кои имаат издолжување: g, j, p, q и y не спаѓаат во групата на преносители на скриени битови и истите не се користат од страна на предложениот метод, со оглед на тоа дека линијата за подвлекување има пресек со нивното издолжување и обојувањето на линијата со бела боја ќе го обои и овој пресек во нијанса на бела боја, со што издолжувањето на знаците ќе има ефект на прекинатост, кој е многу забележителен за крајните корисници.

3.1..3 Манипулација со границите на параграфите

Параграфот разгледуван како група од реченици кои се поврзани една по друга и при тоа одделен од другите параграфи со знак за премин во нова линија, сам по себе може да има одредени својства во софтверот *Microsoft Word* . Како својство кое е искористено за пренос на скриени битови во предложениот метод, се границите на параграфите. Граници во *Microsoft Word* посебно може да се додаваат на параграфите, реченициите, сликите, табелите, на секоја страна

посебно итн. и секој од овие ентитети може да има лева, десна, горна и долна граница.

Исто како и линиите за подвлекување од претходно, така и границите имаат множество од стилови кои може да се изберат. Постојат 24 различни стила на граници и со експерименти е дефинирано дека само два стилови се забележителни за човековото око (стиловите „*wdLineStyleEmboss3D*“ и „*wdLineStyleEngrave3D*“). Со користење на преостанатите 22 можни стила, може да се дефинира мапа со 16 вредности, на начин што секој стил од тие 22 вредности, одговара на некоја од вредностите помеѓу 0000 – 1111. Со тоа, стилот на една граница е потенцијален преносител на 4 скриени бита.

Слично како и претходно, со цел границата на параграфот да помине незабележително, на истата се задава бела боја, на начин што таа се аплицира само на границата и таа се синхронизира со позадината и не е видлива за човековото око. Слично како и во претходниот метод, се дефинираат најблиските 16 вредности на бела боја (*R,G,B* >= 253) и со тоа изборот на боја на границата на параграфот исто така станува преносител на четири бита.

Во разгледаниот оптимален случај, секоја граница на параграф е потенцијален преносител на 8 бита (4+4). Преку експерименти е дефинирано дека со цел да помине незабележително за крајниот корисник, манипулацијата со границите може да се изврши само на левата и на десната граница т.е. користењето на горната и долната граница создава голем простор помеѓу линиите во параграфот, што е многу забележливо од страна на крајниот корисник. Од овде произлегува дека капацитетот на основната верзија на методот директно зависи од бројот на параграфи во документот (што и не е многу голем капацитет) и секој параграф е потенцијален преносител на 16 бита (8 од левата граница и 8 од десната граница).

При предлагањето на алтернативи на главниот метод, слично како и претходно се разгледува користењето на најблиските 128 нијанси на бела боја од достапните 216 вредности (*R,G,B* >= 249). Тие се прифатливи со оглед на тоа дека не се забележливи за човековото око, но истите покрај тоа што овозможуваат поголем капацитет за 11 бита (7+4) по граница т.е. 22 бита по параграф, истите носат поголем ризик за нивно откривање.

Друга алтернатива е проширувањето на капацитетот преку стилот на границите. Секој стил на граници има дополнителни својства меѓу кои спаѓа и ширината на границата. Со разгледување на ширините на достапните 22 стила кои можат да се користат за границите, дефинирано е дека:

- за 13 стила на границите се достапни 9 различни ширини на границите;
- за 2 стила на границите се достапни 6 различни ширини на границите;
- за 3 стила на границите се достапни 5 различни ширини на границите;
- еден стил има 8 различни ширини на границите;
- еден стил има 2 различни ширини на границите;
- два од стиловите имаат достапно по само една ширина на границите.

Со соодветни комбинации на стиловите на границите со нивните ширини, може да се дојде до тоа да стиловите да можат да бидат преносители, исто така по 7 бита. На пример, со избор само на стиловите кои имаат достапни по 8 или 9 ширини на границите (тоа се 14 достапни стилови), кај истите самиот стил врши пренос на четири бита по граница, додека изборот на ширина овозможува пренос на дополнителни 3 бита по граница т.е. вкупно 7 бита по стил (4+3). Со избор на стиловите кои имаат достапни по 5 или 6 ширини на границите (тоа се 5 достапни стилови), кај истите самиот стил врши пренос на четири бита по граница, додека изборот на ширина овозможува пренос на дополнителни 2 бита по граница т.е. вкупно 6 бита по стил (4+2). Исто како и алтернативата со боите, и алтернативата со стиловите го зголемува капацитетот за пренос, но во исто време го зголемува и ризикот да методот биде полесно забележан од страна на крајниот корисник.

3.1..4 Манипулација со границите на реченициte

Слично како и на параграфите, така и на реченициte може посебно да се дефинираат границите. Користењето на границите на реченициte овозможува поголем капацитет за пренос на скриените пораки (со оглед на тоа дека дефинитивно тие се повеќе бројни), но самите граници се сместени директно во текстот, па истите многу полесно може да бидат забележливи за разлика од параграфите. И овој метод се базира на манипулации со левата и со десната граница, но истите бараат поголема внимателност со оглед на тоа дека колку границата е поголема/поширока толку повеќе празниот простор помеѓу местото

каде што завршува една реченица е подалечно од местото каде што почнува нова реченица.

При додавање на граници на реченициите, се достапни 16 видови на стилови, но токму поради нивната големина / ширина, преку експерименти е дефинирано дека само 8 од нив се соодветни за користење на начин што нема да бидат препознаени. При тоа, на секој од овие 8 достапни стила треба да биде користена најмалата можна ширина. Стиловите што се погодни за да се употребат се: „*wdLineStyleDash-Dot*“, „*wdLineStyleDashDotDot*“, „*wdLineStyleDashLargeGap*“, „*wdLineStyleDashSmallGap*“, „*wdLineStyleDot*“, „*wdLineStyleInset*“, „*wdLineStyleOutset*“ и „*wdLineStyleSingle*“.

Со цел границата на реченицата да помине незабележително, на истата се задава бела боја, на начин што таа се аплицира само на границата и со тоа се синхронизира со позадината и не е видлива за човековото око. И овде се дефинираат најблиските 16 вредности на бела боја (*R,G,B* >= 253) и со тоа изборот на боја на границата на реченицата станува преносител на четири бита. Секоја граница на реченица е преносител на 7 бита (3 бита од стилот и четири бита од бојата).

Предложена алтернатива е само онаа за проширување на капацитетот преку изборот на боја (*R,G,B* >= 249) со што секоја граница би овозможила пренос на 10 бита (3 бита од стилот и 7 бита од бојата). При манипулацијата со границите на реченициите, се препорачува да се користи само едната граница (или левата или десната), бидејќи кога се користат и двете, на местото каде што една реченица завршува а друга почнува доаѓа до преклопување на десната граница од првата и левата граница од втората, со што празниот простор помеѓу реченициите дополнително се зголемува, што е индикатор кој може многу лесно да биде препознаен од страна на крајниот корисник.

3.2 Предложени методи за стегананализа

Предложени се повеќе методи за вршење на стеганализа на *Microsoft Word* документи, кои се дел од категоријата на методи базирани на формати. По спроведувањето на стеганализата, резултатите се добиваат во форма на статистики, каде што се прикажани фреквенциите на својствата кои се карактеристични за конкретниот метод кои се анализира.

3.2..1 Стеганализа на отворените методи

Методот врши анализа за тоа дали во рамките на еден документ е содржана скриена порака вгнездена со отворениот метод за реченици или со отворениот метод за зборови т.е. дали има додавање на дополнителни и непотребни празни места по завршувањето на секоја реченица и по секој збор.

При анализата на стеганографскиот отворен метод за реченици, се пресметува вкупниот број на реченици во документот и бројот на реченици кои имаат повеќе од едно празно место по завршувањето. Реченициte кои имаат повеќе од едно празно место, се всушност потенцијални ентитети за пренос на скриени битови, со оглед на тоа доколку стеганографскиот отворен метод е применет. Токму тие дополнителни празни места по реченициte ќе бидат преносителите. Со тоа, како резултат на крајниот корисник кој ја врши анализата, се прикажува информацијата колку вкупно реченици има во документот и колкав број од нив се потенцијални преносители на скриени битови.

При анализата на стеганографскиот отворен метод за зборови, се пребројува вкупниот број на зборови во документот и бројот на зборови кои имаат повеќе од едно празно место по самиот збор. Зборовите по кои следува повеќе од едно празно место, се всушност потенцијални ентитети за пренос на скриени битови, па слично како и кај реченициte, како резултат на крајниот корисник кој ја врши анализата, се прикажува информацијата колку вкупно зборови има во документот и колкав број од нив се потенцијални преносители на скриени битови.

Одлуката за тоа дали во документот е применет некој од стеганографските отворени методи, останува на крајниот корисник. Со разгледување на резултатите од анализите, тој самиот проценува за веројатноста на тоа дали бројот на потенцијалните преносители е доволен за да се смета дека во документот постои скриена порака, вгнездена со отворен метод.

3.2..2 Стеганализа на методи кои вршат манипулација на невидливите знаци

Како што беше опишано претходно, постојат поголем број на методи кои ги искористуваат специјалните невидливи знаци со цел пренос на скриени битови.

Предложената анализа ги чита сите знаци од документот и ја брои фреквенцијата на секој од знаците:

- *Right remark (U+200E);*
- *Left remark (U+200F);*
- *Zero width joiner (U+200D);*
- *Zero width non-joiner (U+200C);*
- *Zero-Width Character (U+200B).*

Како резултат од анализата се прикажуваат кодовите на пронајдените невидливи специјални знаци и за секој од нив е дадено колку пати истиот се појавил во рамките на документот.

Одлуката за тоа дали во документот е применет некој од стеганографските методи кој ги искористува овие невидливи знаци, останува на крајниот корисник. Со разгледување на резултатите од анализите, тој самиот проценува за веројатноста на тоа дали бројот на потенцијалните преносители е доволен за да се смета дека во документот постои скриена порака, вгнездена со некој метод кои вметнува специјални невидливи знаци.

Во категоријата на методи за манипулација на невидливите знаци влегува и методот со кој се менува бојата на невидливите знаци (празните места – *space*, празните места – *tab* и знаците за премин во нова линија), со што самите *RGB* вредности претставуваат преносители на скриените битови.

Анализата за конкретниот метод се врши на начин што се читаат сите знаци од документот и се брои вкупниот број на невидливи знаци (трите типови на празни места споменати погоре). Дополнително, се брои бројот на премини на боите на невидливите знаци од една боја во друга, при споредба на два соседни невидливи знака т.е. се споредуваат боите на секој пар соседни невидливи знаци и доколку тие имаат различни бои, тоа се класифицира како премин од една боја во друга.

На пример, во еден секојдневен *Microsoft Word* документ (книга, барање, изјава, скрипта итн.) бојата на сите знаци е иста (најчесто црна) и при читањето на еден ваков документ, нема да биде пронајден ниту еден премин од една боја во друга, што само по себе е многу јасен индикатор за тоа дека соодветниот стеганографски метод не е користен во документот.

3.2..3 Стеганализа на методот базиран на типови на фонтови

Анализата за тоа дали врз документот е применет стеганографскиот метод базиран на типови на фонтови, со менување на првите големи букви на зборовите, се врши така што се наоѓа вкупниот број на зборови во документот кои започнуваат со голема буква и се наоѓа вкупниот број на премини од еден фонт во друг. Слично како и претходно така и овде, се споредуваат сите соседни потенцијални ентитети за пренос т.е. во овој случај сите зборови кои започнуваат со голема прва буква и се споредуваат фонтовите на правите букви од соседните зборови. Доколку двете букви имаат различни фонтови, тогаш парот се вбројува како премин од еден фонт во друг.

Дополнително, се бројат и различните фонтови кои се користени како фонтови на првите букви, па корисникот ќе има подобра претстава за тоа дали при промените на фонтовите постојано се повторуваат само одредени фонтови, или пак постојано станува збор за нови различни фонтови.

На крајот, како резултат се добива вкупниот број на зборови кои започнуваат со прва голема буква, заедно со бројот на премини на тие првите големи букви од еден фонт во друг и со бројот на различни фонтови кои се користени во првите големи букви на зборовите.

Кај *Microsoft Word* документите кои секојдневно се користат, бројот на вакви премини од еден фонт во друг скоро и да не постои (или е минимален) па малите вредности на овој излезен параметар се доволен индикатор за констатација дека методот не се користи во документот.

3.2..4 Стеганализа на методот за скалирање на знаците

За предложениот стеганографски метод кој користи скалирање на знаците, предложен е и соодветен метод за негова стеганализа. Со вчитување на сите знаци од документот, се проверува својството за скалирање на секој од нив.

Се креира мапа на начин што секоја вредност која што е прочитана како скалирана вредност, се сместува како главна вредност и потоа се брои колку пати истата таа вредност всушност се појавила како скалирана вредност на некој карактер. Како резултат од методот, се прикажуваат скалираните вредности и за секоја вредност се прикажува нејзината фреквенција.

Со оглед на тоа дека преддефинирана вредност за скалирањето во *Microsoft Word* е 100%, оваа вредност се маркира како „*default*“ вредност и при приказот на резултатите за истата како име се користи таа референца, а за сите други вредности се прикажуваат тие самите.

Доколку стеганографскиот метод за скалирање не е применет, очекувано е најчесто да има појава само на „*default*“ вредноста и таа се појавува онолку пати колку што има знаци во документот. Прифатливо е да има варијации и мали фреквенции и на некои други вредности, но појавата на голема фреквенција на вредности кои се блиски до вредноста 100%, е сериозен индикатор за тоа дека во документот се применува скалирање на знаците, со цел вгнездување на скриени пораки.

3.2..5 Стеганализа на методот за подвлекување на знаците

За предложениот стеганографски метод кој користи подвлекување на знаците, предложен е и соодветен метод за негова стеганализа. Со вчитување на сите знаци од документот, се проверува својството за подвлекување на секој од нив.

За секој подвлечен знак, се чита стилот кој е користен за линијата за подвлекување, како и бојата која што е користена врз самата линија. Стилот и бојата се претвораат во текстуални вредности и нивните имиња се поврзуваат со знакот „-“ со што се добива името на користената комбинација стил/боја за линијата користена при подвлекувањето на знаците. Ваквата комбинација се сместува во мапа, каде што се чуваат и се бројат сите комбинации во рамките на документот. Бидејќи името на стилот или бојата може да е многу долго и нечитливо, на секоја комбинација се додава кратко име со соодветен индекс на крајот на името („*comb 1*“, „*comb 2*“, „*comb 3*“ …).

Како резултат од методот, се прикажуваат кратките имињата на пронајдените комбинации и за секоја од нив се прикажува соодветната фреквенција. Со оглед на тоа дека комбинацијата на стилот „*wdUnderlineNone*“ со бојата „*automatic*“ е комбинација која што ќе биде препознаена за знаците кои што не се подвлечени, истата се маркира и се именува како „*default*“ со цел корисникот да биде свесен дека таа може многу често да се појавува, без притоа

да значи дека е преносител на скриени битови. Поради нејзината честа појава, на корисникот со ова му се дава до знаење дека може да ја игнорира.

Доколку во резултатот од анализата се лоцира појава само на комбинацијата „*default*“, без разлика на нејзината фреквенција, овој резултат е индикатор дека во документот со сигурност може да се констатира дека соодветниот метод не е користен за вгнездување на скриени пораки. За останатите случаи, се остава на корисникот самиот да ја процени веројатноста, врз основа на добиените резултати.

3.2..6 Стеганализа на методот за манипулација со границите на параграфите

За предложениот стеганографски метод кој врши манипулации со границите на параграфите, предложен е и соодветен метод за негова стеганализа. Со вчитување на сите параграфи од документот, се проверува дали истите имаат лева или десна граница и доколку имаат некоја од нив, се преминува кон проверка на својствата на границите што ги имаат.

За секоја постоечка лева или десна граница на параграф, се чита стилот кој е користен за линијата на самата граница, како и бојата која што е користена за истата. Слично како и претходно, стилот и бојата се претвораат во текстуални вредности и нивните имиња се поврзуваат со знакот „-“ со што се добива името на користената комбинација стил/боја за соодветната граница. Ваквата комбинација се сместува во мапа, каде што се чуваат и се бројат сите комбинации во рамките на документот. Бидејќи името на стилот или бојата може да е многу долго и нечитливо, на секоја комбинација се додава кратко име со соодветен индекс на крајот на името („*comb 1*“, „*comb 2*“, „*comb 3*“ …).

За разлика од претходниот метод за стеганализа, во овој случај се креираат две мапи, така што во една мапа се бројат комбинациите од левите граници на параграфите, а од десната страна се бројат комбинациите од десните граници на параграфите.

Како резултат од методот, се прикажуваат две посебни листи, кои се соодветно обележани дека ги претставуваат комбинациите кои се појавуваат во левата / десната граница соодветно. Листите ги содржат кратките имињата на пронајдените комбинации и за секоја од нив се прикажува соодветната

фреквенција. Со оглед на тоа дека комбинацијата на стилот „*wdLineStyleNone*“ со бојата „*automatic*“ е комбинација која што ќе биде препознаена во случај и воопшто да нема поставено граници на параграфот, истата се маркира и се именува како „*default*“ со цел корисникот да биде свесен дека е очекувано таа многу често да се појавува, без притоа да значи дека е преносител на скриени битови. Со тоа на корисникот му се дава до знаење дека истата може да ја занемари при анализа на добиените резултати.

3.2..7 Стеганализа на методот за манипулација со границите на реченициите

Овој метод е идентичен како и предложениот метод за стеганализа на границите на параграфите, само што овде станува збор за границите на реченициите. Тој одговара за детектирање на горе предложениот стеганографски метод кој врши манипулации со границите на реченициите. Се вчитуваат сите реченици од документот, се проверува дали истите имаат лева или десна граница и доколку имаат некоја од нив, се преминува кон проверка на својствата на границите што ги имаат.

За секоја постоечка лева или десна граница на реченица, се чита стилот кој е користен за линијата на самата граница, како и бојата која што е користена за истата. Исто како и претходно, стилот и бојата се претвораат во текстуални вредности и нивните имиња се поврзуваат со знакот „-“ со што се добива името на користената комбинација стил/боја за соодветната граница. Ваквата комбинација се сместува во мапа, каде што се чуваат и се бројат сите комбинации во рамките на документот. Бидејќи името на стилот или бојата може да е многу долго и нечитливо, на секоја комбинација се додава кратко име со соодветен индекс на крајот на името („*comb 1*“, „*comb 2*“, „*comb 3*“ …).

Исто како и кај параграфите, и овде се креираат две мапи, така што во една мапа се бројат комбинациите од левите граници на реченициите, а од десната страна се бројат комбинациите од десните граници на реченициите.

Како резултат од методот, се прикажуваат две посебни листи, кои се соодветно обележани дека ги претставуваат комбинациите кој се појавуваат во левата/десната граница соодветно. Листите ги содржат кратките имињата на пронајдените комбинации и за секоја од нив се прикажува соодветната фреквенција. Со оглед на тоа дека комбинацијата на стилот „*wdLineStyleNone*“

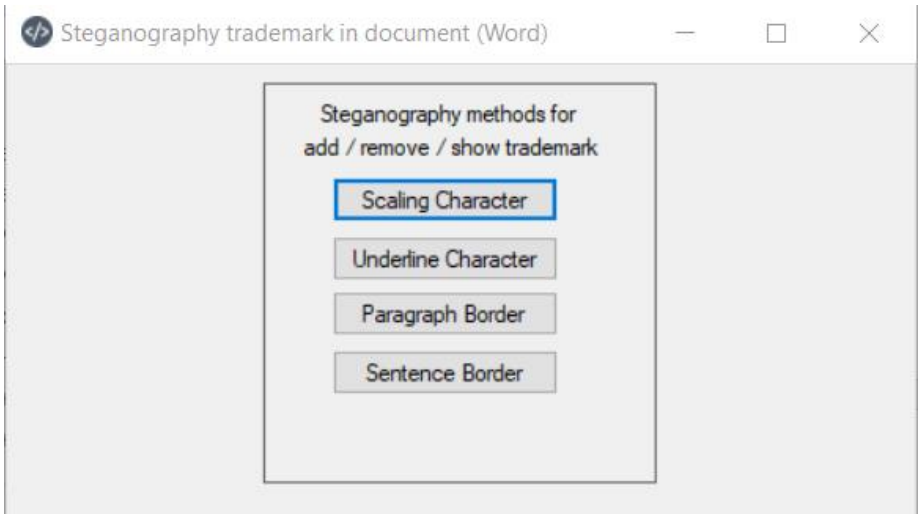
со бојата „*automatic*“ е комбинација која што ќе биде препознаена во случај и воопшто да нема поставено граници на реченицата, истата се маркира и се именува како „*default*“ со цел корисникот да биде свесен дека е очекувано таа многу често да се појавува, без притоа да значи дека е преносител на скриени битови. Со тоа на корисникот му се дава до знаење дека истата може да ја игнорира при анализа на добиените резултати.

4 Дискусија

4.1..1 Имплементација на методите за стеганографија

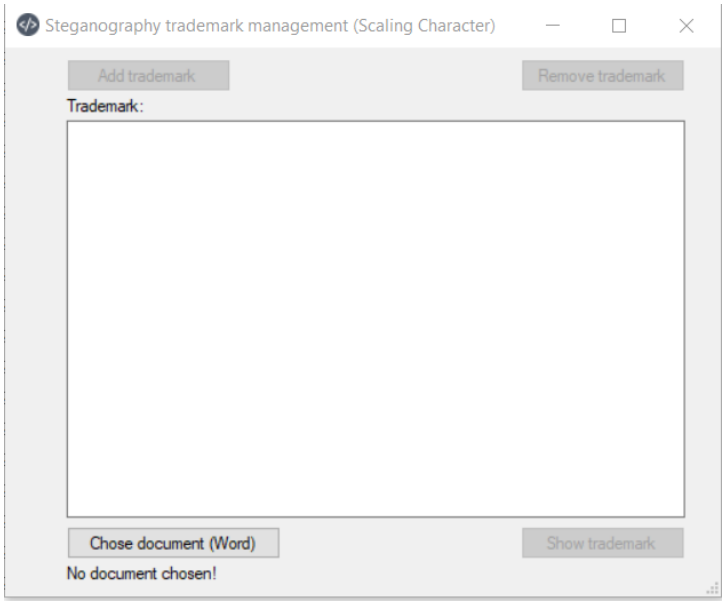
Предложените стеганографски методи се имплементирани во програмскиот јазик *Visual C#*, преку алатка која ја нарековме „*MSWordST*“ и која овозможува вгнездување на скриени пораки во даден документ и подоцна нивно соодветно вчитување. Под претпоставка дека и испраќачот и примачот ја имаат истата алатка, тие многу лесно ќе можат да комуницираат со користење на алатката, со оглед на тоа дека таа програмски ги имплементира предложените методи.

На почетниот екран на алатката, корисник избира еден од предложените методи што ќе сака да го користи.



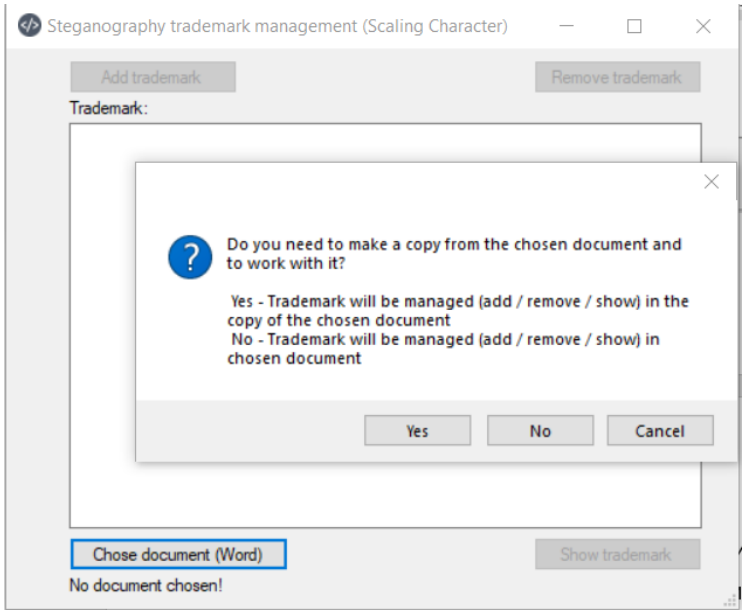
Слика 14: Почетен екран на алатката со методите за стеганографија

Изборот на кој било метод од главниот екран го води корисникот кон идентичен екран со истите опции, а разликата е само во тоа што во позадина се извршува различен метод за криење на пораката.



Слика 15: Почетен екран при изборот на методот за скалирање на знаци

За да може да продолжи понатаму, корисникот треба да прикачи *Microsoft Word* документ во којшто сака да вметне скриена порака, преку копчето „*Chose document (Word)*“, по што се појавува соодветна порака и избор од страна на корисникот за тоа дали директно да се работи врз избраниот документ (со што неговата оригинална содржина ќе биде презапишана) или пак сака да се направи соодветна копија и да се продолжи да се работи со истата.



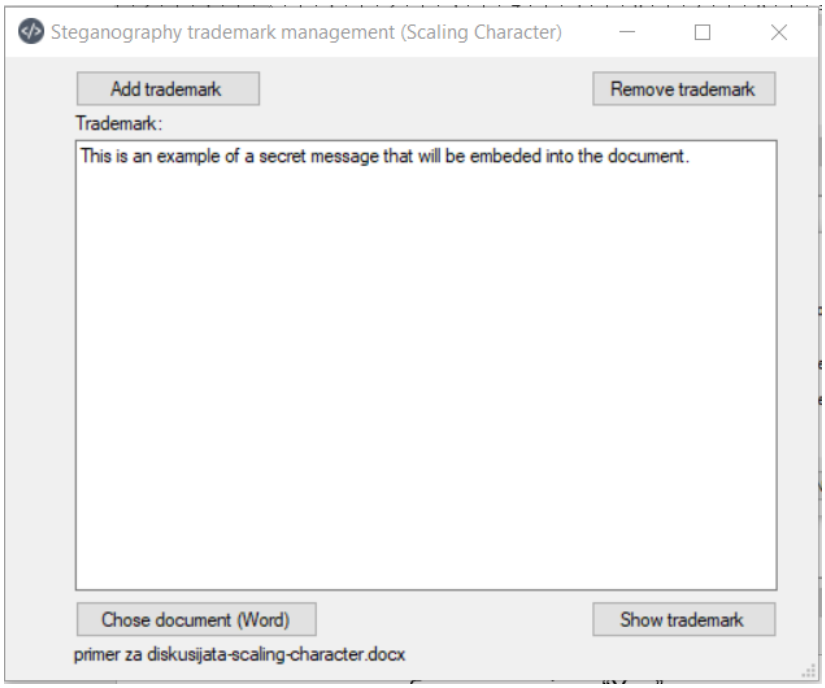
Слика 16: Порака која се појавува при избор на MS Word документ во кој треба да се вгнезди скриена порака

- со избор на опцијата „Yes“ – се креира копија и се продолжува да се работи со неа;
- со избор на опцијата „No“ – се работи директно со оригиналот и содржината се презапишува.

При изборот на опцијата за креирање на копија, новиот документ се креира во истиот фолдер каде што е и оригиналот и при тоа го задржува истото име како и оригиналот и на крајот се додава името на методот кој е користен:

- доколку се користи методот за скалирање на знаци, по оригиналниото име се додава текстот „*-scaling-character*“;
- доколку се користи методот за подвлекување на знаци, по оригиналниото име се додава текстот „*-underline-character*“;
- доколку се користи методот за манипулација со границите на параграфите, по оригиналниото име се додава текстот „*-paragraph-border*“;
- доколку се користи методот за манипулација со границите на реченициите, по оригиналниото име се додава текстот „*-sentence-border*“.

Името е прикажано во лабелата на екранот, а во конкретниов пример, името на оригиналниот документ е „*primer za diskusija*“.

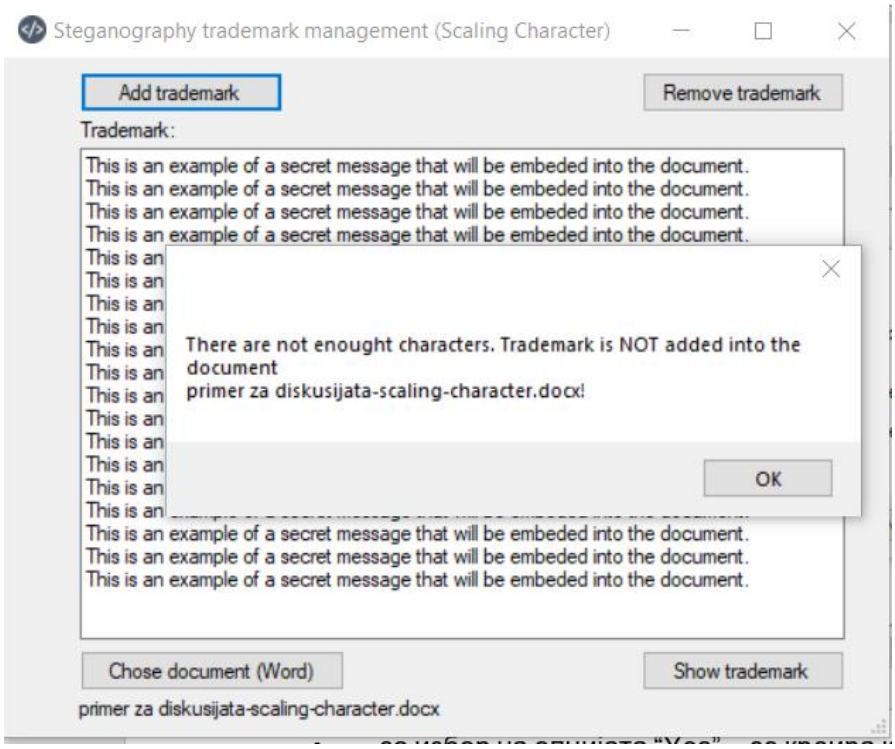


Слика 17: Вчитување на документ и пишување на скриена порака што треба да се вгнезди

По вчитувањето на документот, сите копчиња стануваат овозможени и корисникот може да пристапи кон пишување на пораката која што сака скришно да ја вгнезди во документот.

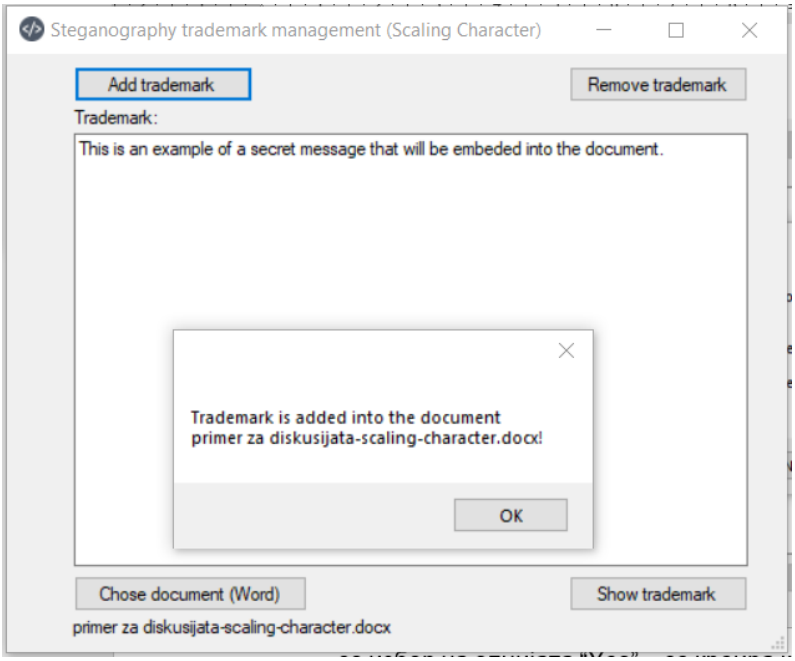
По пишувањето на текстот на скриената порака, стеганографскиот метод за вгнездување започнува да се извршува со клик на копчето „*Add trademark*“, по што се добива соодветна порака за статусот на тоа дали пораката е успешно внесена или не.

Доколку бројот на соодветните ентитети за пренос во рамките на оригиналниот документ, не одговара на капацитетот на избраниот метод т.е. доколку скриената порака е подолга од максималниот капацитет кој документот го поддржува, тогаш корисникот ќе добие соодветна порака со известување дека пораката е премногу долга и дека документот не е соодветен за неа.



Слика 18: Порака за неуспешно вгнездување, поради многу долга скриена порака т.е. недоволен капацитет на документот

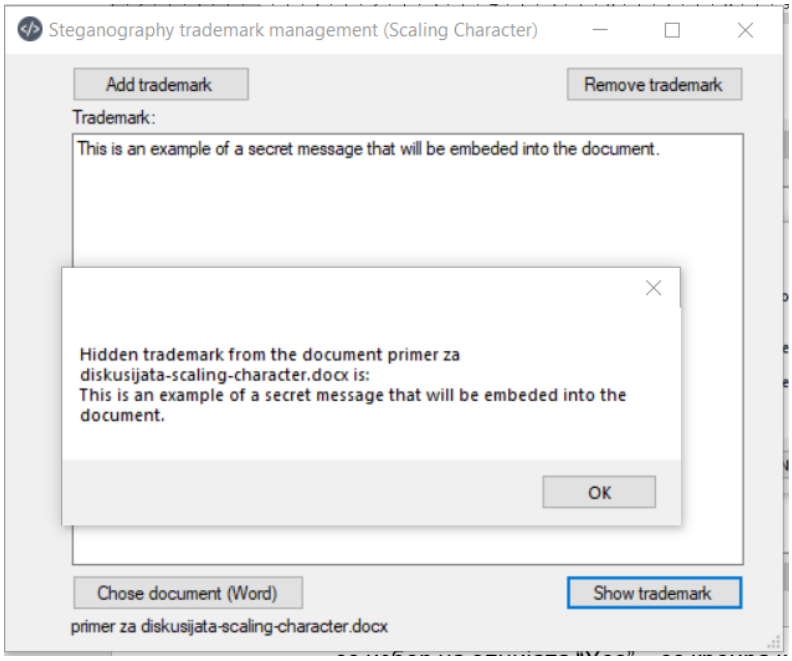
При успешно вгнездување, се добива соодветна порака за истото.



Слика 19: Порака за успешно вгнездување на скриена порака во документ

Како што беше споменато и претходно, истата алатка се користи и на страната на примачот. Откако го добива документот кој е генериран со процесот опишан погоре, тој го избира соодветниот метод во алатката и го вчитува документот за кој веќе знае дека содржи скриена порака.

Пораката се декодира со клик на копчето „*Show trademark*“.

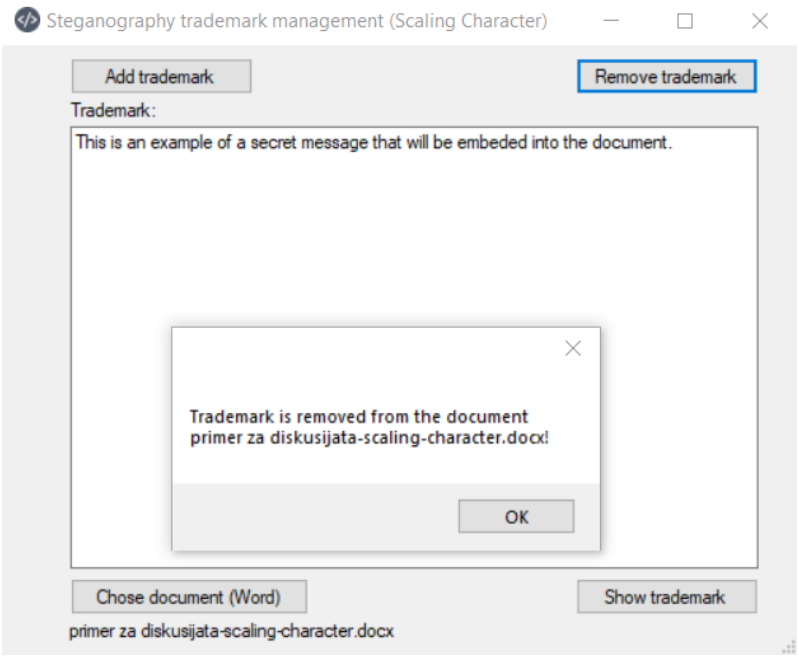


Слика 20: Декодирање на веќе вгнездена скриена порака во документ

Доколку корисникот сака да продолжи да го препраќа документот без притоа документот да има скриена порака, таа може да се отстрани со клик на копчето „*Remove trademark*“.

Ова сценарио за отстранување на скриената порака може да се искористи доколку примачот сака да го проследува документот без скриена порака, или пак доколку испраќачот и примачот имаат договор за комуникација само преку одредени документи. На страната на примачот, тој може да ја декодира пораката од документот кој го примил, потоа пораката може да ја отстрани преку алатката и самиот да внесе нова порака во истиот тој документ, со што сега тој ќе ја добие улогата на испраќач.

На тој начин, документот може цело време да останува ист и само пораката која ја носи да се менува (под претпоставка дека корисниците ја избираат опцијата за презапишување на документот).

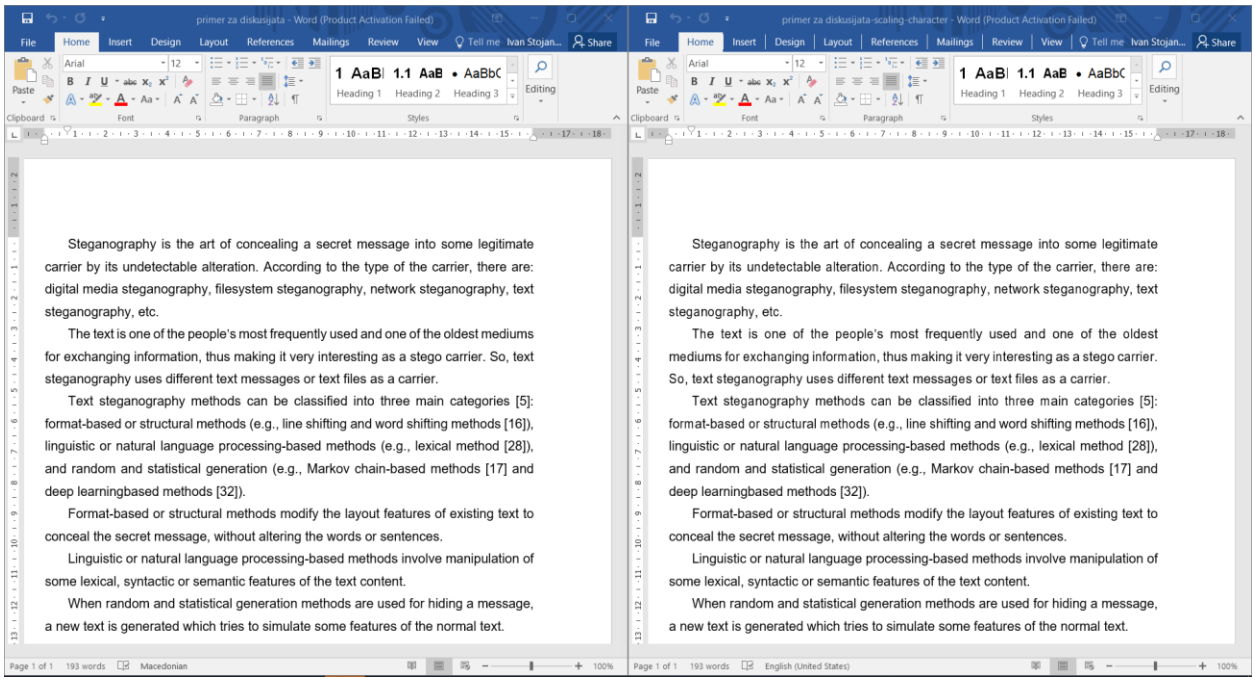


Слика 21: Отстранување на веќе вгнездена скриена порака од документ

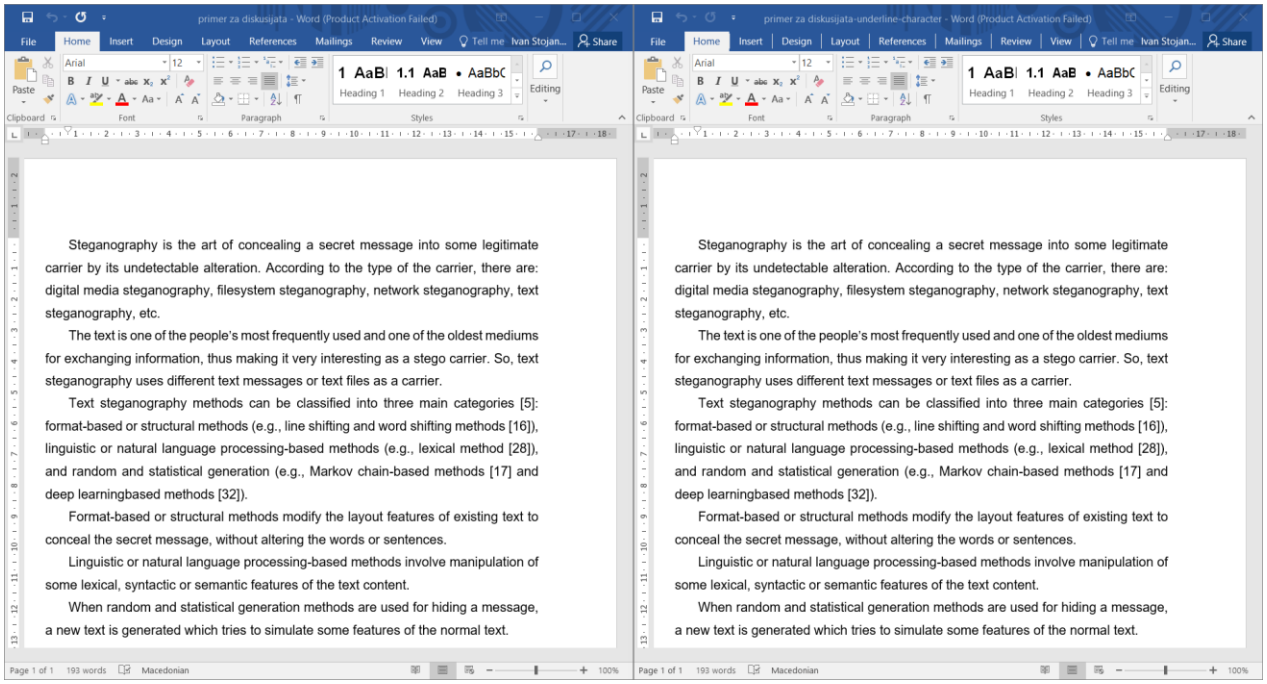
Како што беше споменато и претходно, сите четири предложени стеганографски методи ја користат истата форма т.е. истиот изглед како сликите од примерите погоре и само начинот на вгнездување на пораките е различен. Единствена разлика е тоа што името на методот кој се користи во даден момент, е прикажано на крајот на насловот од формата (даден во заграда).

Во продолжение се дадени примери за споредба на тоа како документите изгледаат пред и после вгнездувањето на скриена порака, за секој од предложените четири методи. Како што може да се забележи во продолжение, под претпоставка дека документите го имаат „најчестиот“ изглед кои генерално *Microsoft Word* документите го имаат, постоењето на променети својства на одредени ентитети (знаци, зборови, реченици, параграфи и сл.) ќе помине незабележливо.

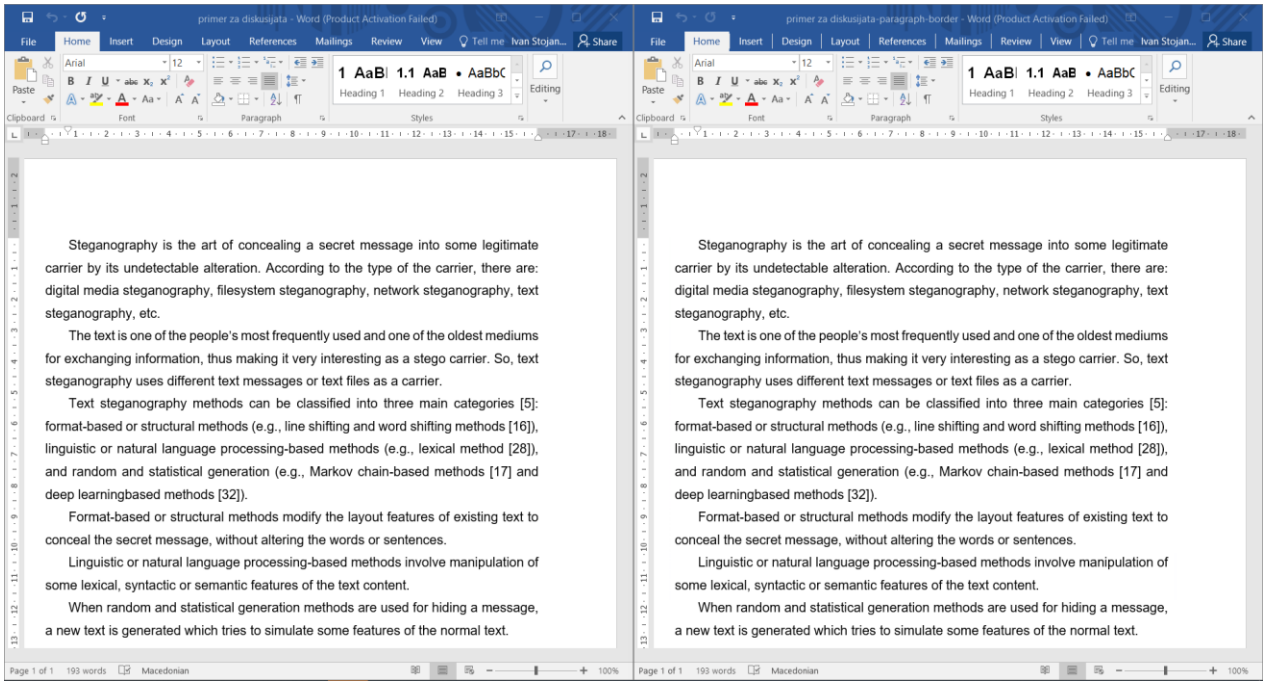
Една од претпоставките на методите е дека документот ќе има бела боја на позадината на страниците, поради и што својствата на ентитетите кои се манипулираат се најчесто зададени во бела боја. Доколку бојата на страниците не е бела, извршените промени врз својствата на ентитетите (линијата за подвлекување и границите) нема да поминат незабележително, со оглед на тоа дека користените нијанси на бела боја најверојатно ќе бидат видливи. Доколку испраќачот / примачот ја знаат бојата на позадината на документот, тие секогаш може да ги променат нијансите на користените нијанси, со цел да ги прилагодат примените да поминат незабележително.



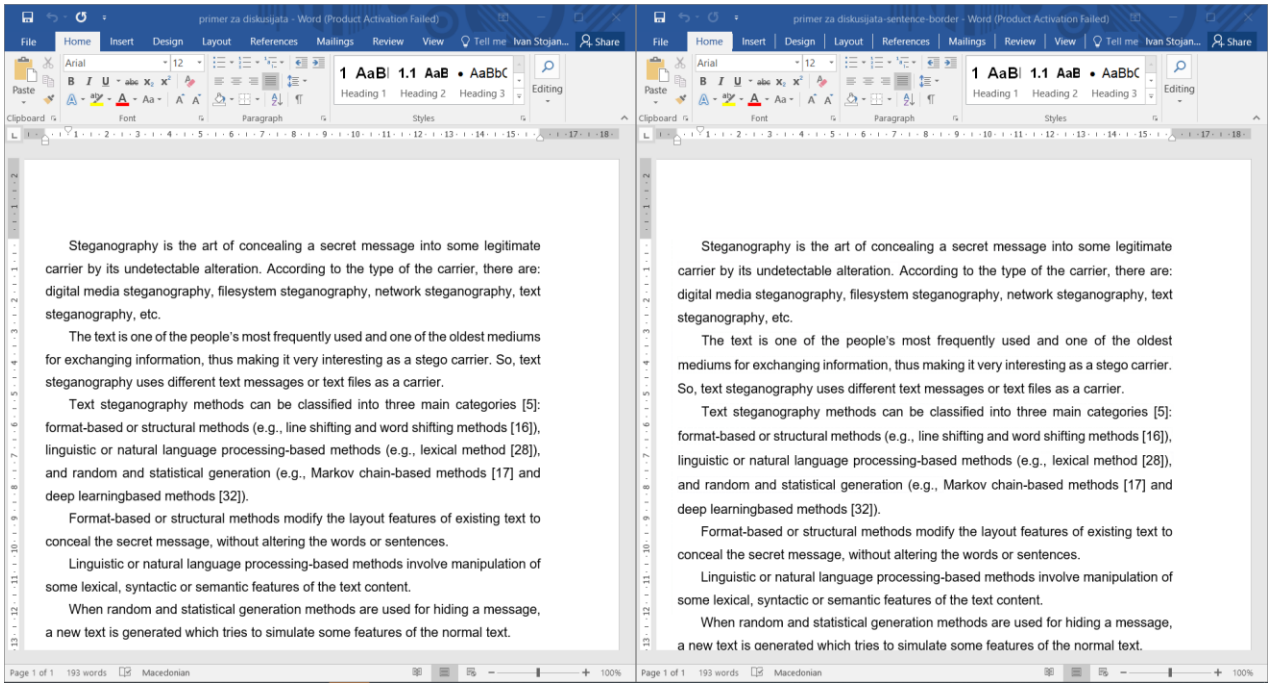
Слика 22: Оригинален документ спореден со документ кои содржи скриена порака со предложениот метод за скалирање на знаците



Слика 23: Оригинален документ спореден со документ кои содржи скриена порака со предложениот метод за подвлекување на знаците



Слика 24: Оригинален документ спореден со документ кои содржи скриена порака со предложениот метод за манипулација со границите на параграфите

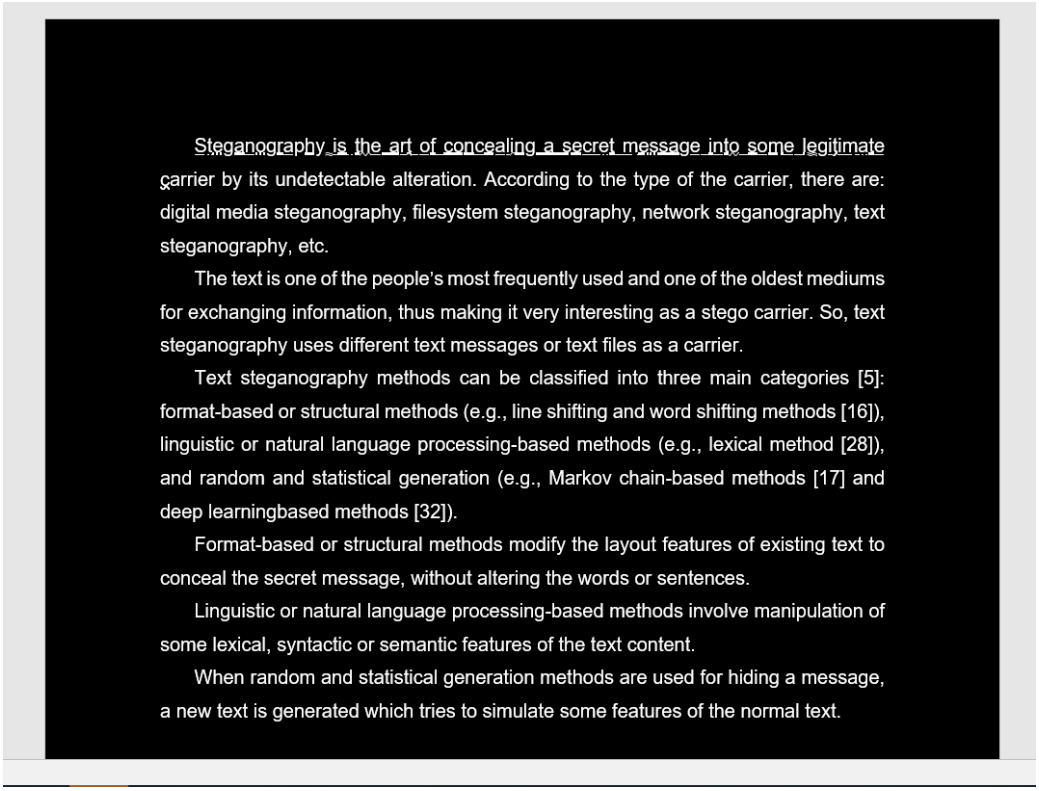


Слика 25: Оригинален документ спореден со документ кои содржи скриена порака со предложениот метод за манипулација со границите на речениците

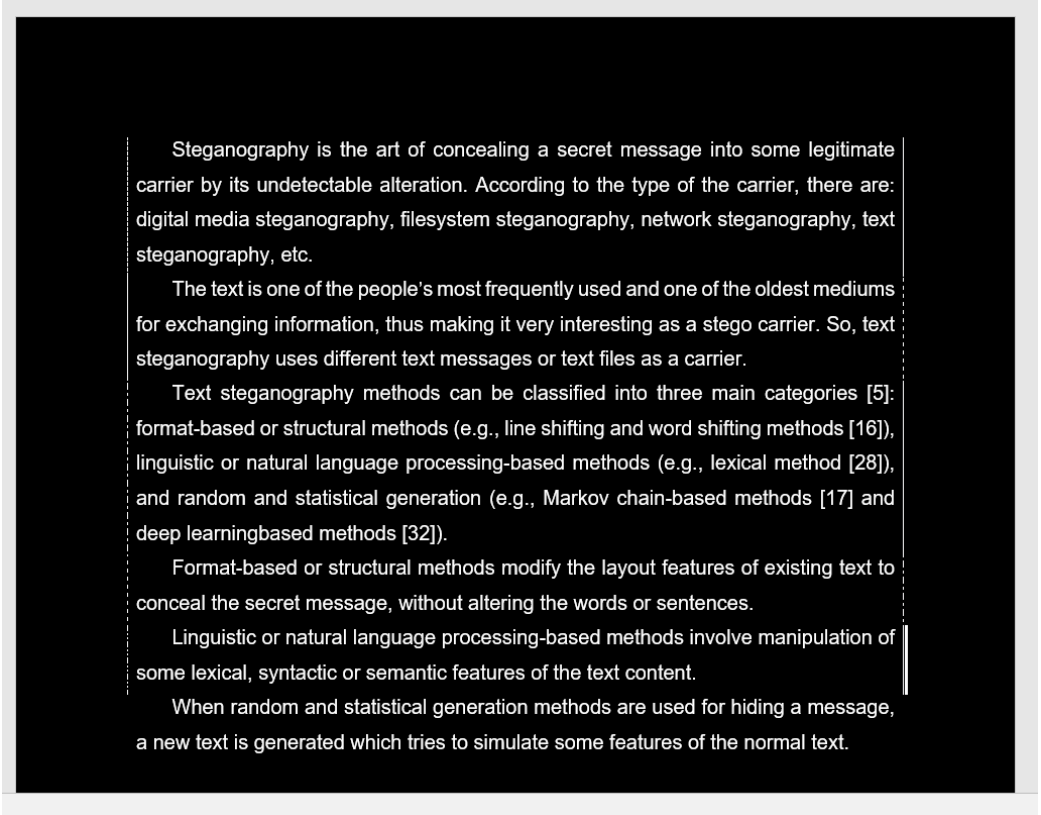
Со цел за полесно да се сфати на кој начин всушност работи методот и кои својства тој ги менува, во продолжение се дадени истите документи од споредбите погоре, со тоа што е променета бојата на позадината во црна боја, со цел да бидат видливи сите оние својства кои се додадени и прикриени со различни нијанси на бела боја.

Кај методот кој врши подвлекување на знаците, на *Слика 26* се гледа како првите знаци се подвлечени со различни типови на линии т.е. со различни стилови на самите линии за подвлекување. Со оглед на тоа што овој метод има поголем капацитет од останатите предложени методи, од самата слика се гледа како за пренос на скриената порака се доволни помал број на знаци од вкупниот број на знаци во документот.

Кај методот кој врши манипулација со границите на параграфите, на *Слика 27* се гледа како од страната на параграфите се присутни граници кои при тоа не навлегуваат во содржината на текстот и поради тоа поминуваат незабележително. Секој параграф има различни типови на граници т.е. граници со различни стилови на самите линии.



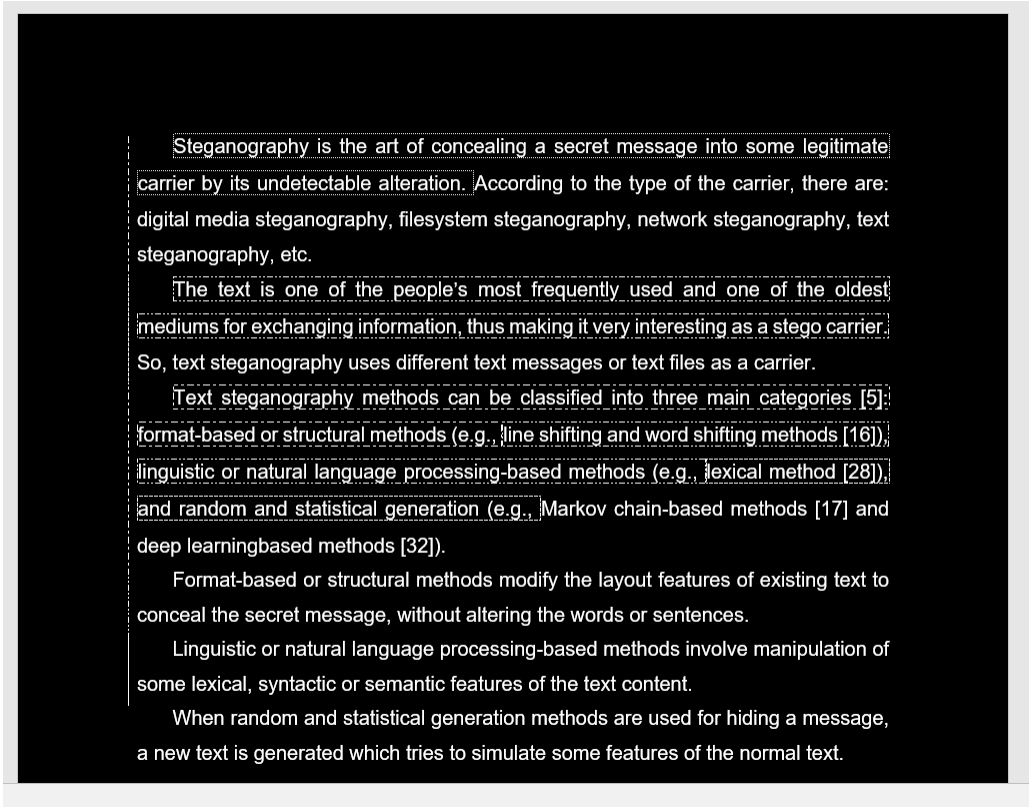
Слика 26: Приказ на промените на својствата во документ кој содржи скриена порака со предложениот метод за подвлекување на знаците



Слика 27: Приказ на промените на својствата во документ кој содржи скриена порака со предложениот метод за манипулација со границите на параграфите

Кај методот кој врши манипулација со границите на реченициите, се гледа како од страната на реченициите се присутни граници кои при тоа навлегуваат во содржината на текстот и поради тоа има опасност да бидат полесно забележителни од претходните предложени методи. Секоја реченица има различни типови на граници т.е. граници со различни стилови на самите линии.

Од приказот за реченициите се гледа како методот е применет само на левата граница на реченициите, поради опасноста од тоа да биде забележан доколку се применува на двете граници истовремено.



Слика 28: Приказ на промените на својствата во документ кој содржи скриена порака со предложениот метод за манипулација со границите на реченициите

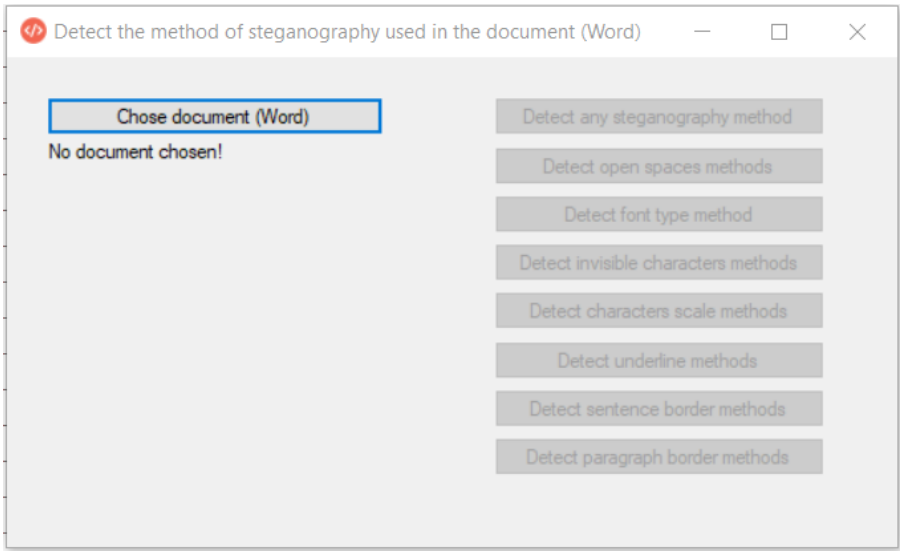
4.1..2 Имплементација на методите за стеганализа

Предложените методи за стеганализа се имплементирани во програмскиот јазик *Visual C#*, преку алатка која ја нарековме „*MSWordSST*“[1] и таа овозможува анализирање на даден документ со цел приказ на одредени статистики кои се карактеристични за одредени стеганографски методи.

[1] https://github.com/ivan0071/SteganalysisApp

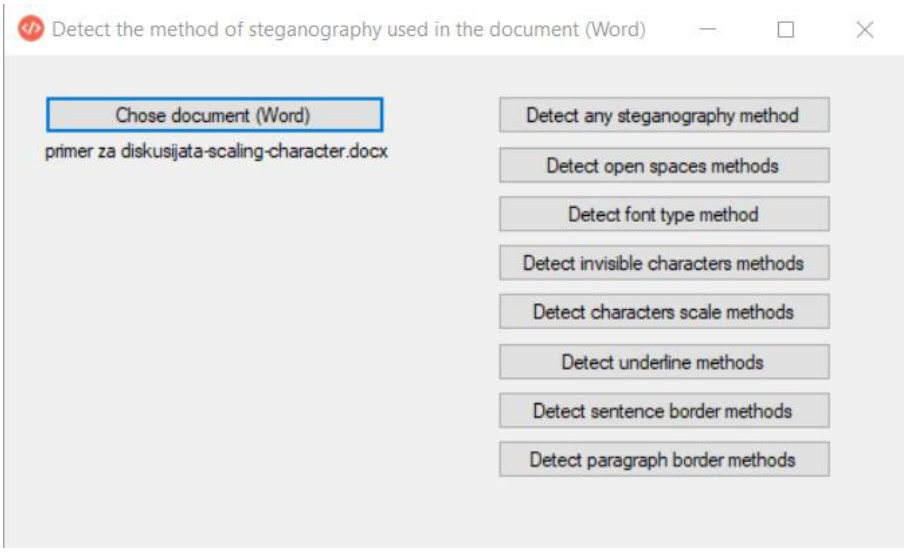
Резултатите од извршената стеганализа која што ја врши алатката, се поврзани само со конкретните предложени методи за стеганализа и истите не даваат корисни информации од аспект на останатите категории на методи, чии што карактеристики не се опфатени во анализата.

На почетниот екран на алатката, корисникот најпрво треба да вчита документ а потоа и да избира со кој од предложените методи ќе сака да ја изврши анализата.



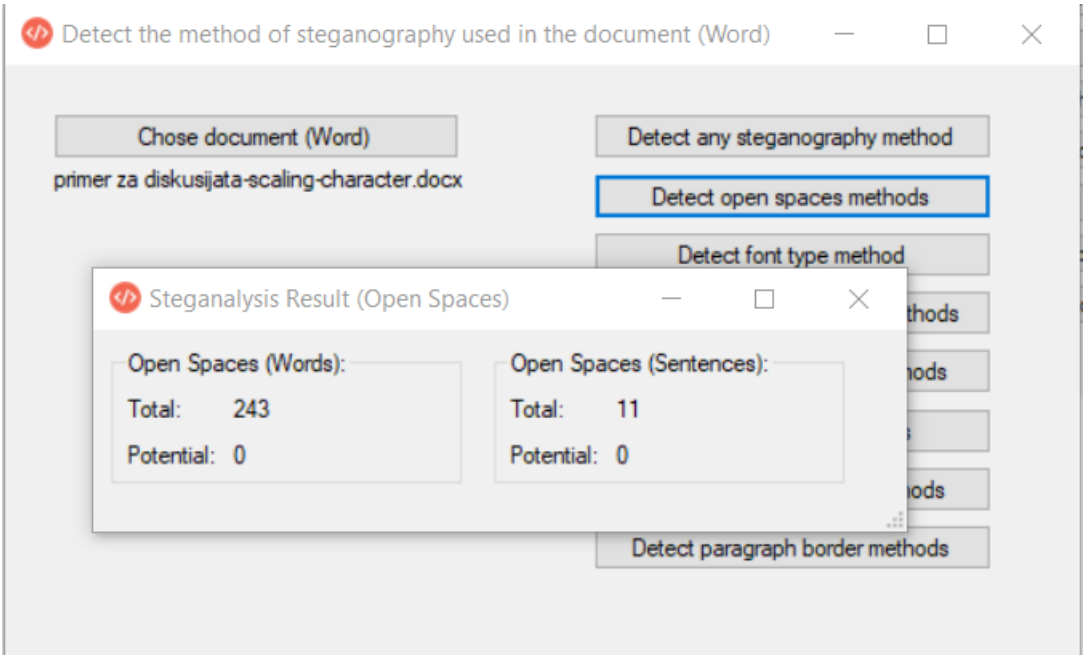
Слика 29: Почетен екран на алатката со методите за стеганализа

Во демонстрациите кои следуваат, се вчитани документите кои беа разгледувани при демонстрацијата на алатката за стеганографските методи.



Слика 30: Вчитување на документ за вршење на стеганализа

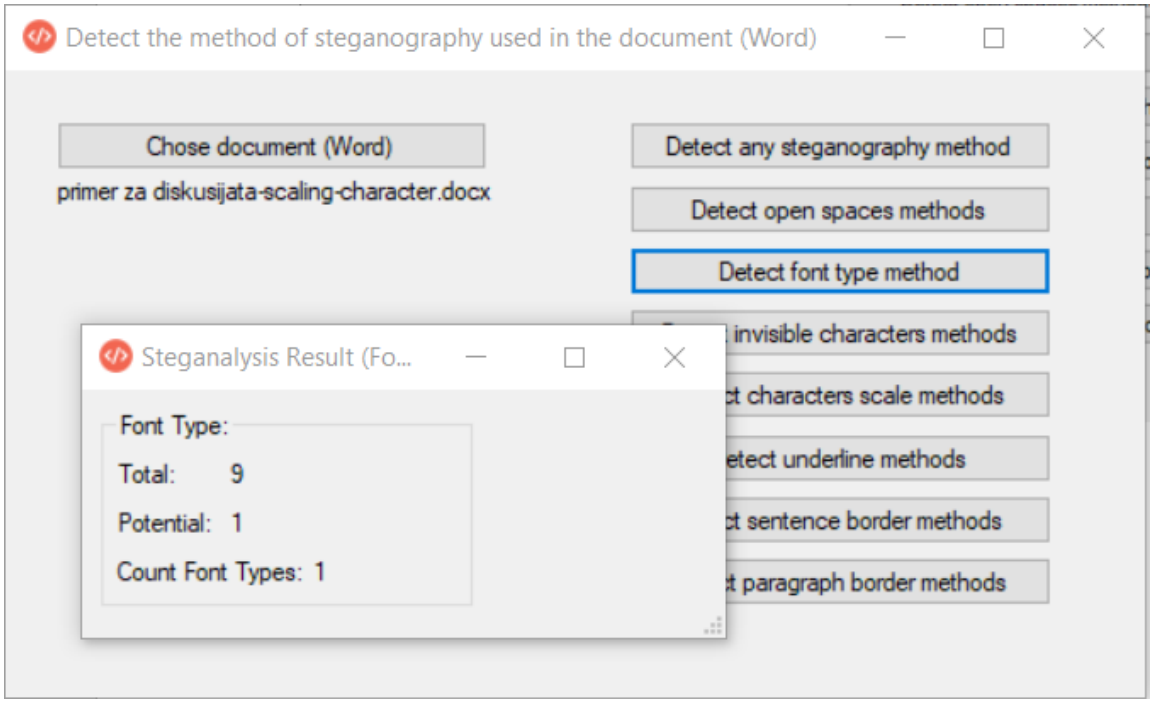
Во продолжение е даден резултатот што го прикажува алатката, во случај кога се врши стеганализа со користење на отворените методи, врз документ на кој е применет предложениот стеганографски метод за скалирање на знаците. Тоа е всушност истиот документ кој е креиран при разгледувањето на алатката за стеганографија.



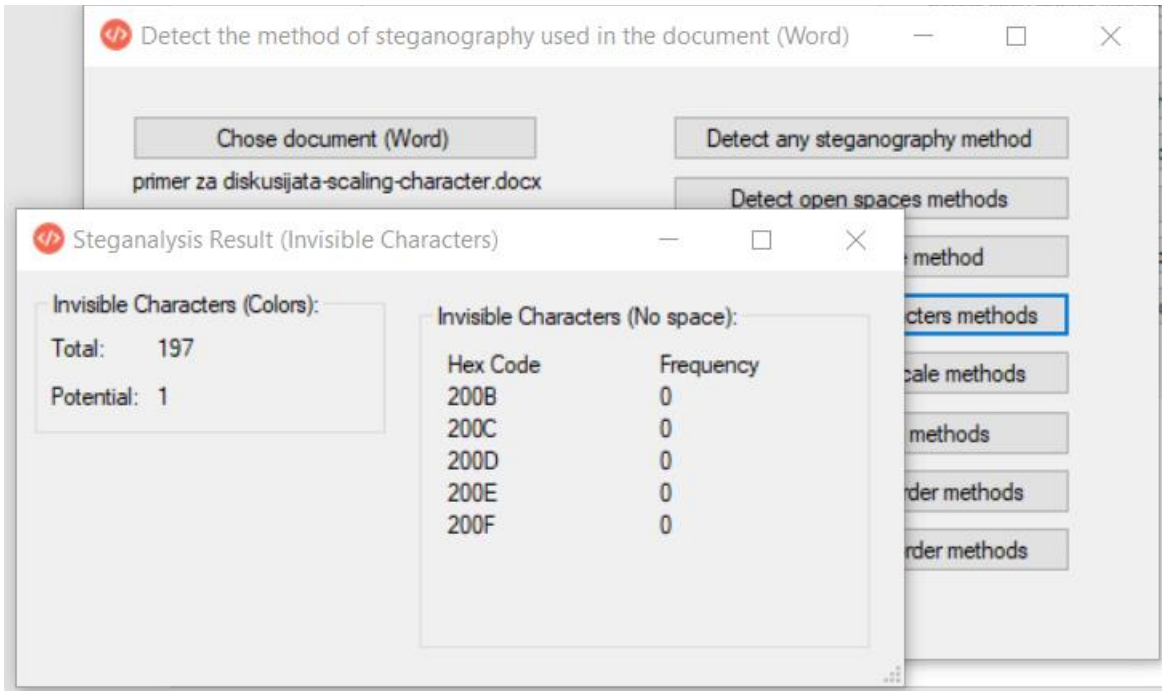
Слика 31: Стеганализа на отворените методи, врз документ на кој е применет стеганографски метод за скалирање на знаци

Како што се гледа од самиот резултат, очигледно е дека ниту еден од анализираните отворени методи не се применети врз вчитаниот документ, со оглед на тоа дека нема пронајдено ниту еден потенцијален ентитет за пренос на скриената порака т.е. нема пронајдено ниту еден збор и ниту една реченица, кои по завршувањето имаат дополнителни празни места.

Во случај кога се врши стеганализа со користење на методот базиран на типовите на фонтовите, врз документ на кој е применет предложениот стеганографски метод за скалирање на знаците, повторно се добива резултат кој покажува дека нема пронајдено фреквенција на промени на фонтовите на големите први букви на зборовите.



Слика 32: Стеганализа на методот базиран на фонтови, врз документ на кој е применет стеганографски метод за скалирање на знаци

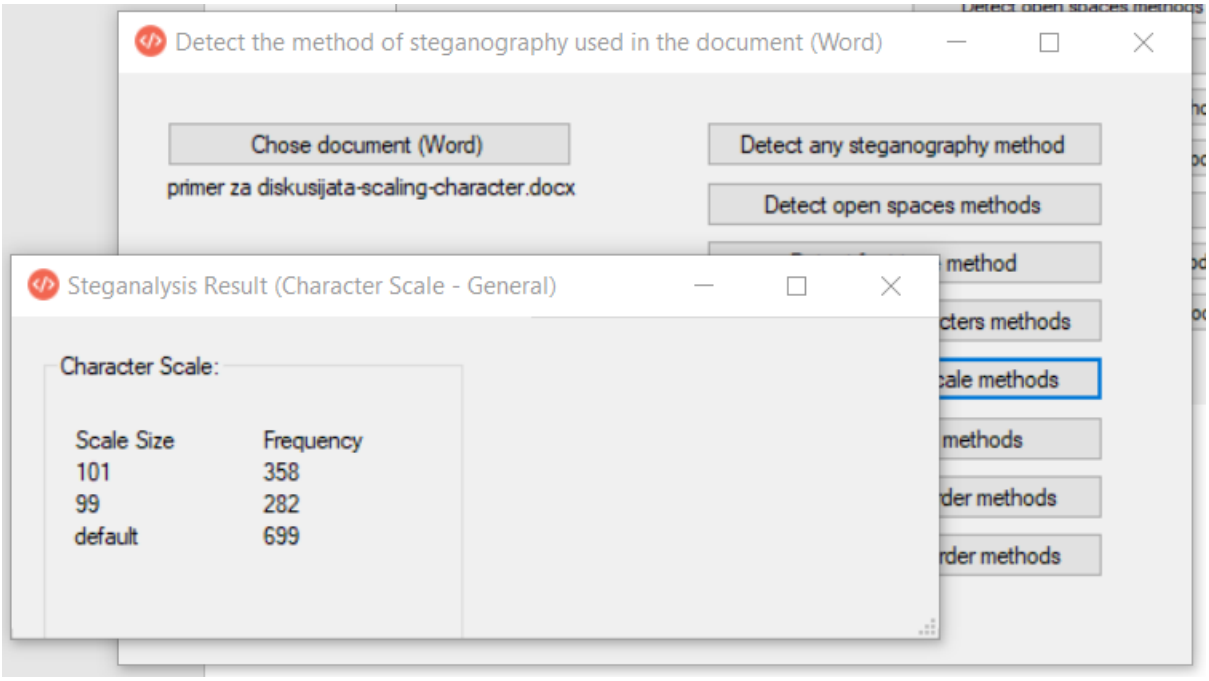


Слика 33: Стеганализа на методите кои вршат манипулација на невидливите знаци, врз документ на кој е применет стеганографски метод за скалирање на знаци

Во случај кога се врши стеганализа со користење на методите кои вршат манипулација на невидливите знаци, врз документ на кој е применет предложениот стеганографски метод за скалирање на знаците, резултат

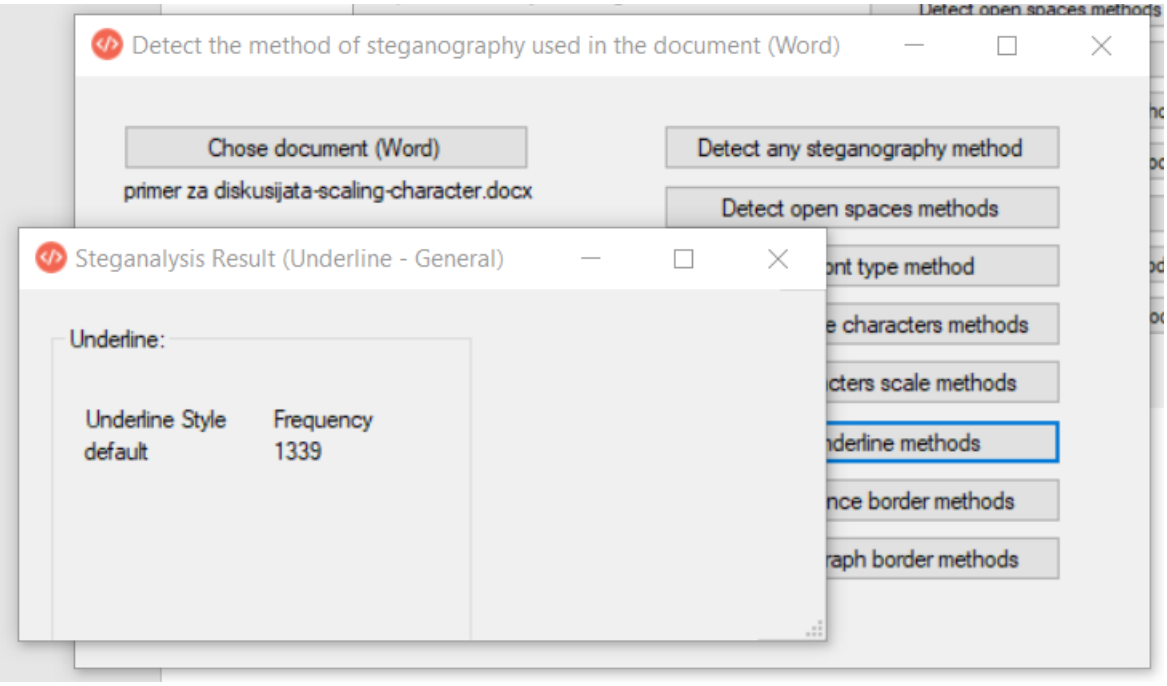
покажува дека нема пронајдено фреквенција на промени на бојата на празните места и покажува дека во документот не се појавуваат специјалните невидливи знаци кои беа предмет на дискусија во соодветните стеганографски методи.

При стеганализа со методот за скалирање на знаците, врз документ на кој е применет истиот тој метод, очекувано е резултатот да прикаже соодветни статистики за одредена фреквенција на вредности кои се блиски до вредноста 100. Во прикажаниот пример во продолжение, се забележуваат поголеми вредности за фреквенциите на вредностите 99 и 101 (вредноста 100 е именувана како „*default*“ вредност) и фактот што нема појава на други вредности, т.е. само најблиските вредности до 100 се јавуваат како вредности за скалирање, е доволен индикатор за тоа да се констатира дека стеганализата успешно извршила потенцијална детекција на стеганографскиот метод за скалирање на знаците.



Слика 34: Стеганализа на методот за скалирање на знаците, врз документ на кој е применет истиот стеганографски метод за скалирање на знаците

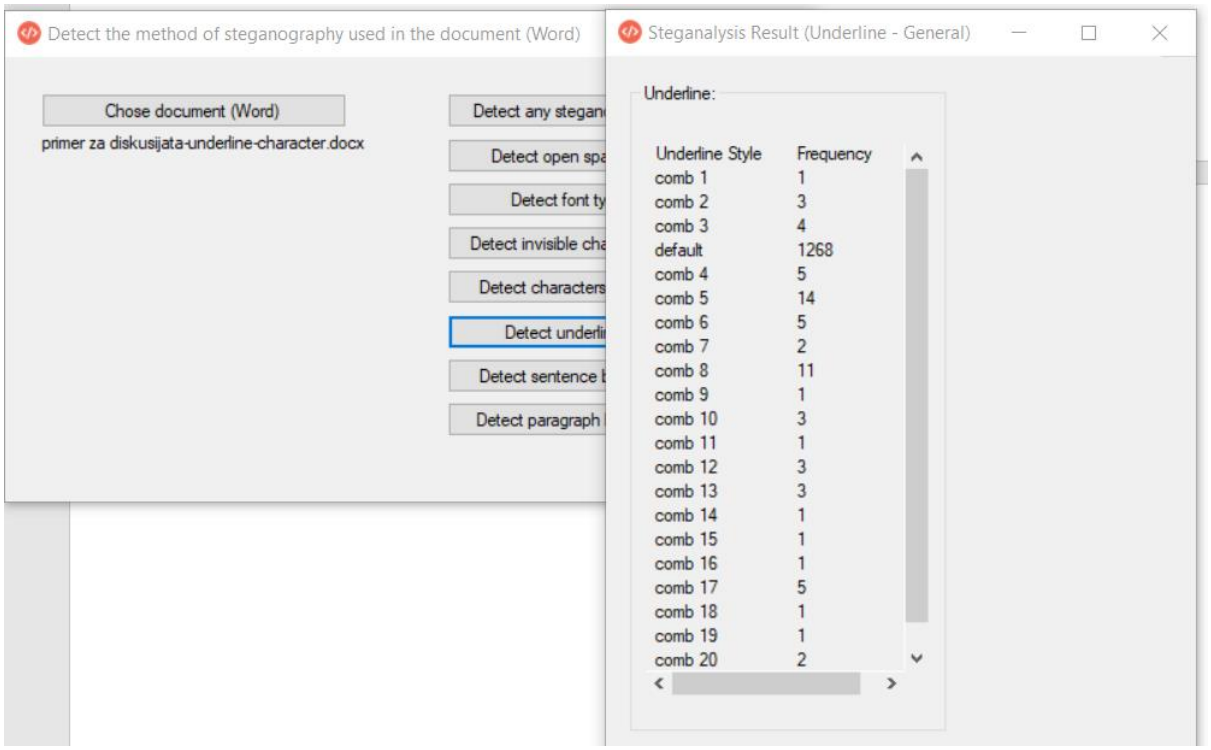
Во случај кога се врши стеганализа со користење на предложениот метод за подвлекување на знаците, врз документ на кој е применет предложениот стеганографски метод за скалирање на знаците, резултат покажува дека во документот нема пронајдено подвлечени знаци кои ги имаат својствата карактеристични за стеганографскиот метод за подвлекување на знаците.



Слика 35: Стеганализа на методот за подвлекување на знаците, врз документ на кој е применет стеганографски метод за скалирање на знаците

Во случај кога се врши стеганализа со користење на предложениот метод за подвлекување на знаците, врз документ на кој е применет предложениот стеганографски метод за подвлекување на знаците, очекувано е резултатот да покажува дека во документот има присуство на голем број на различни комбинации карактеристични за самиот метод кои имаат едно или повеќе појавувања.

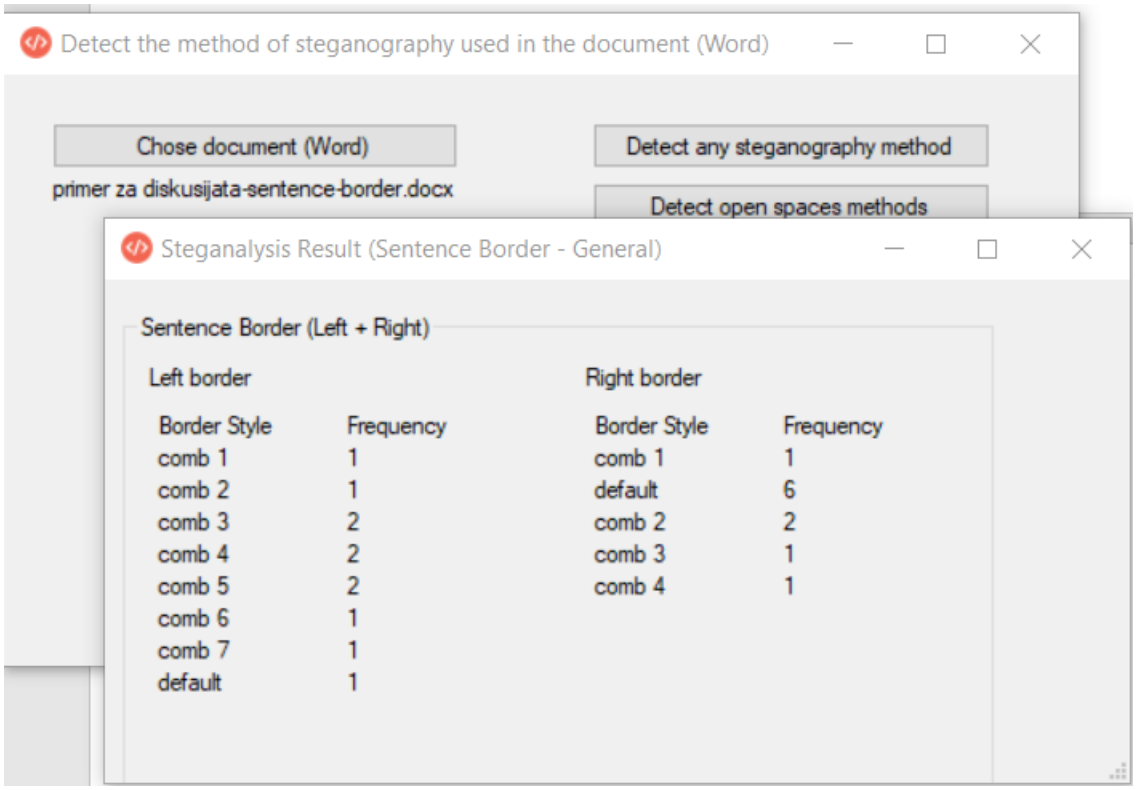
Со оглед на тоа дека резултатот од конкретниот пример покажува присуство на повеќе од 20 комбинации кои ги имаат својствата карактеристични за стеганографскиот метод за подвлекување на знаците, тоа само по себе е доволна индикација за потенцијално присуство на скриена порака, вгнездена со методот за кој е извршена анализата.



Слика 36: Стеганализа на методот за подвлекување на знаците, врз документ на кој е применет истиот стеганографски метод за подвлекување на знаците

Во случај кога се врши стеганализа со користење на предложениот метод за манипулација на границите на речениците, врз документ на кој е применет предложениот стеганографски метод за манипулација на границите на речениците, очекувано е резултатот да покажува дека во документот има присуство на голем број на различни комбинации карактеристични за самиот метод кои имаат едно или повеќе појавувања.

Со оглед на тоа дека резултатот од конкретниот пример покажува присуство на 7 различни комбинации на левата граница и 4 различни комбинации на десната граница кои ги имаат својствата карактеристични за стеганографскиот метод за манипулација на границите на речениците, претставува индикација за потенцијално присуство на скриена порака, вгнездена со методот за кои анализата е извршена.



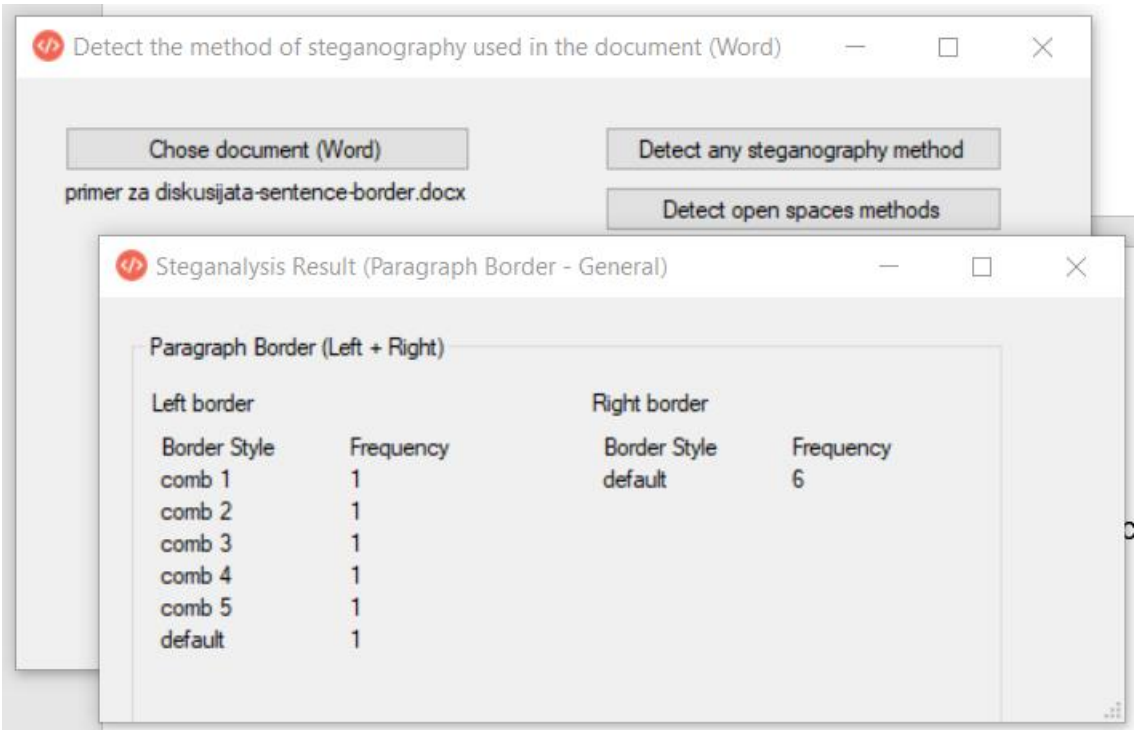
Слика 37: Стеганализа на методот за манипулација со границите на реченициите, врз документ на кој е применет истиот стеганографски метод за манипулација со границите на реченициите

Во случај кога се врши стеганализа со користење на предложениот метод за манипулација на границите на параграфите, врз документ на кој е применет предложениот стеганографски метод за манипулација на границите на реченициите, резултатот може да покажува дека во документот има присуство на голем број на различни комбинации карактеристични за самиот метод кои имаат едно или повеќе појавувања.

Причината за тоа е што двата методи за манипулација на границите (на реченициите и параграфите) се многу слични и поставувањето на лева граница на една реченица, доколку истата таа реченица е прва реченица во параграфот, се преклопува со границата на параграфот.

Во конкретниот пример, документот каде што е вгнездена скриената порака со стеганографскиот метод за манипулација на границите на реченициите има 6 параграфи, па токму затоа резултатите прикажуваат постоење на 6 карактеристични комбинации за левите граници на параграфите. Од ова произлегува дека при анализата на методите со манипулација на границите на

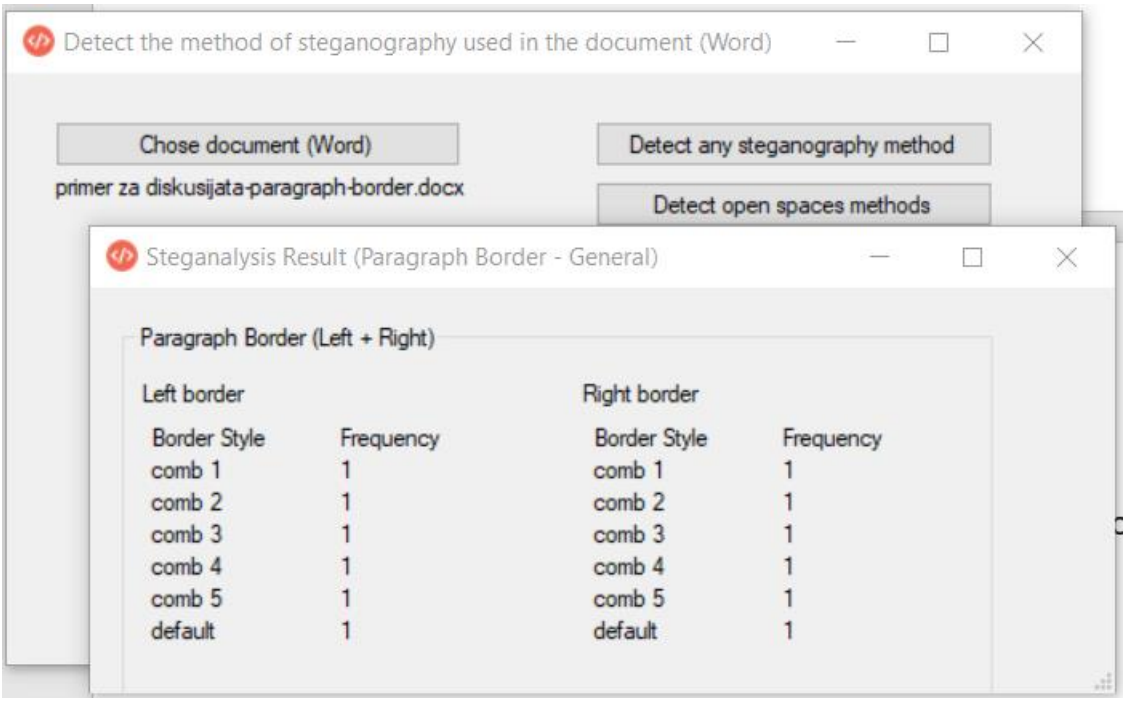
параграфите и реченициите, двата методи меѓу себе може да се преклопуваат со што секогаш може да постои одредено ниво на грешка при анализата.



Слика 38: Стеганализа на методот за манипулација со границите на параграфите, врз документ на кој е применет стеганографски метод за манипулација со границите на реченициите

Во случај кога се врши стеганализа со користење на предложениот метод за манипулација на границите на параграфите, врз документ на кој е применет предложениот стеганографски метод за манипулација на границите на параграфите, очекувано е резултатот да покажува дека во документот има присуство на голем број на различни комбинации карактеристични за самиот метод кои имаат едно или повеќе појавувања.

Со оглед на тоа дека резултатот од конкретниот пример покажува присуство на 5 различни комбинации на левата граница и 5 различни комбинации на десната граница кои ги имаат својствата карактеристични за стеганографскиот метод за манипулација на границите на реченициите, претставува индикација за потенцијално присуство на скриена порака, вгнездена со методот за кои анализата е извршена.

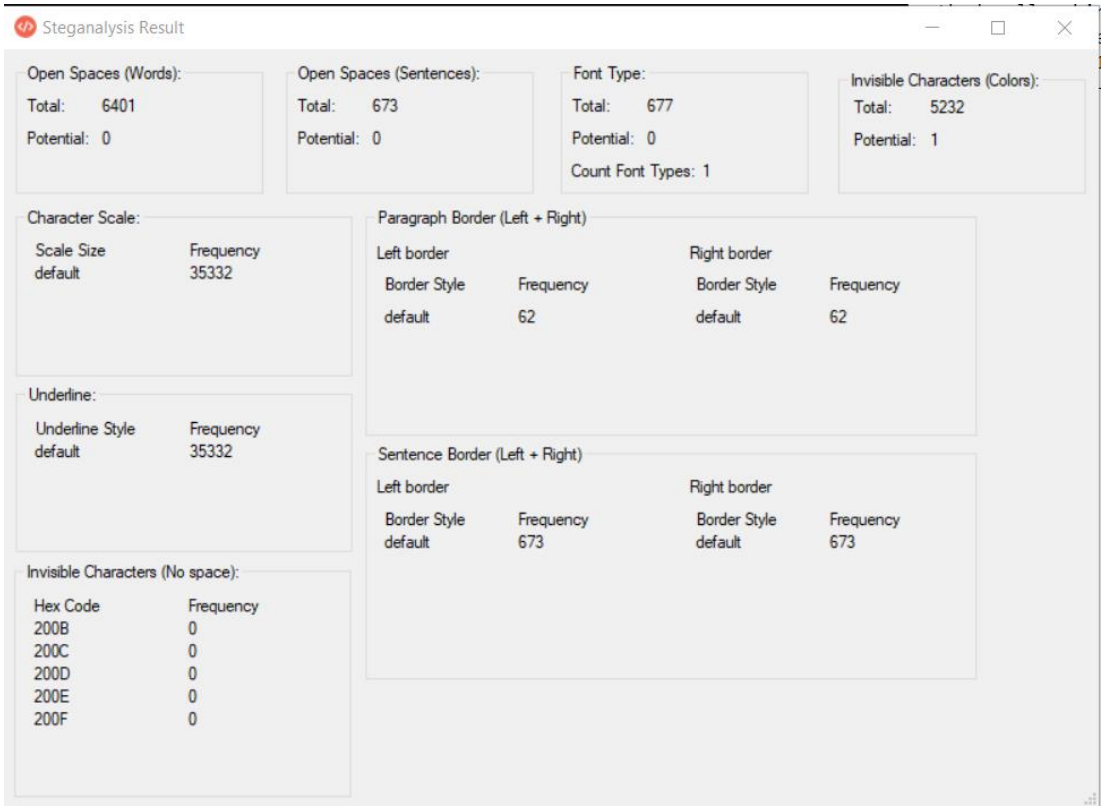


Слика 39: Стеганализа на методот за манипулација со границите на параграфите, врз документ на кој е применет истиот стеганографски метод за манипулација со границите на параграфите

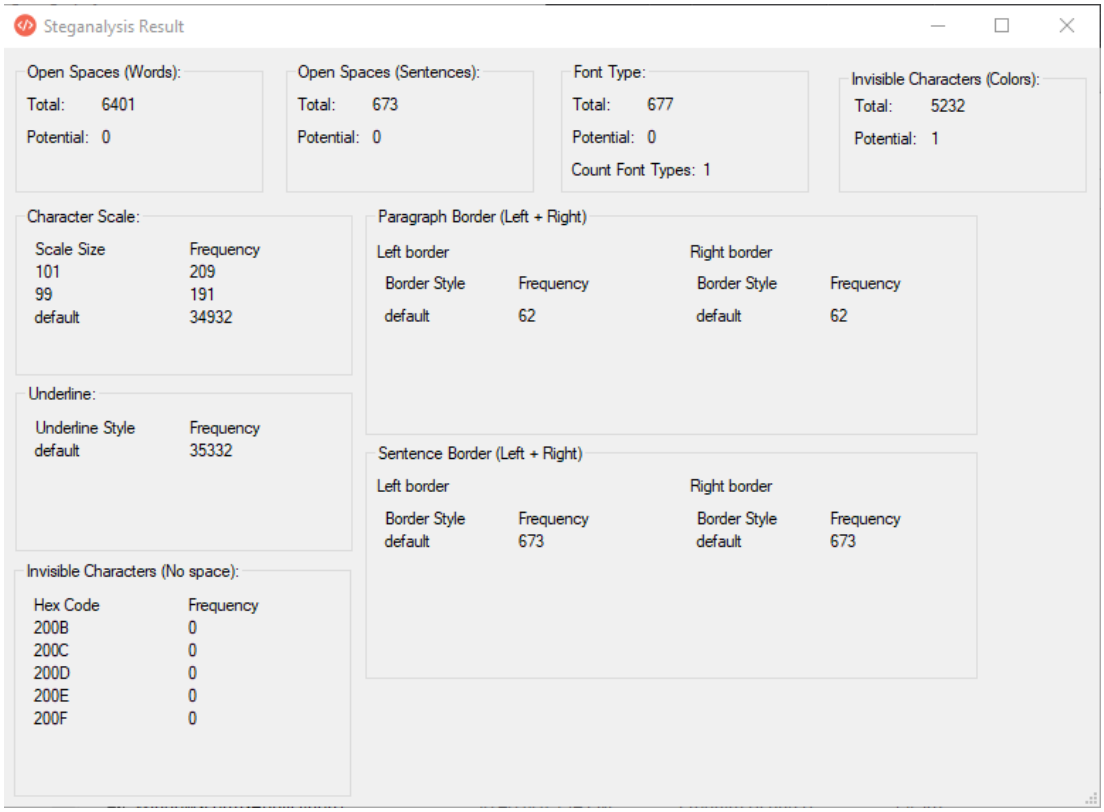
Сепак, поради преклопувањето на двата методи опишано претходно, резултатите од методите за манипулација со границите може да покажуваат присуство на скриената порака, но истите не може да определат за сигурно дали како ентитети за пренос се јавуваат речениците или пак параграфите.

Дополнителна опција на алатката е комбиниран преглед кои ги вклучува резултатите од сите предложени методи за стеганализата. На тој начин, со вчитување на еден документ, се започнува детална стеганализа која што ги испитува сите предложени методи и дава соодветни резултати комбинирани во еден екрански приказ.

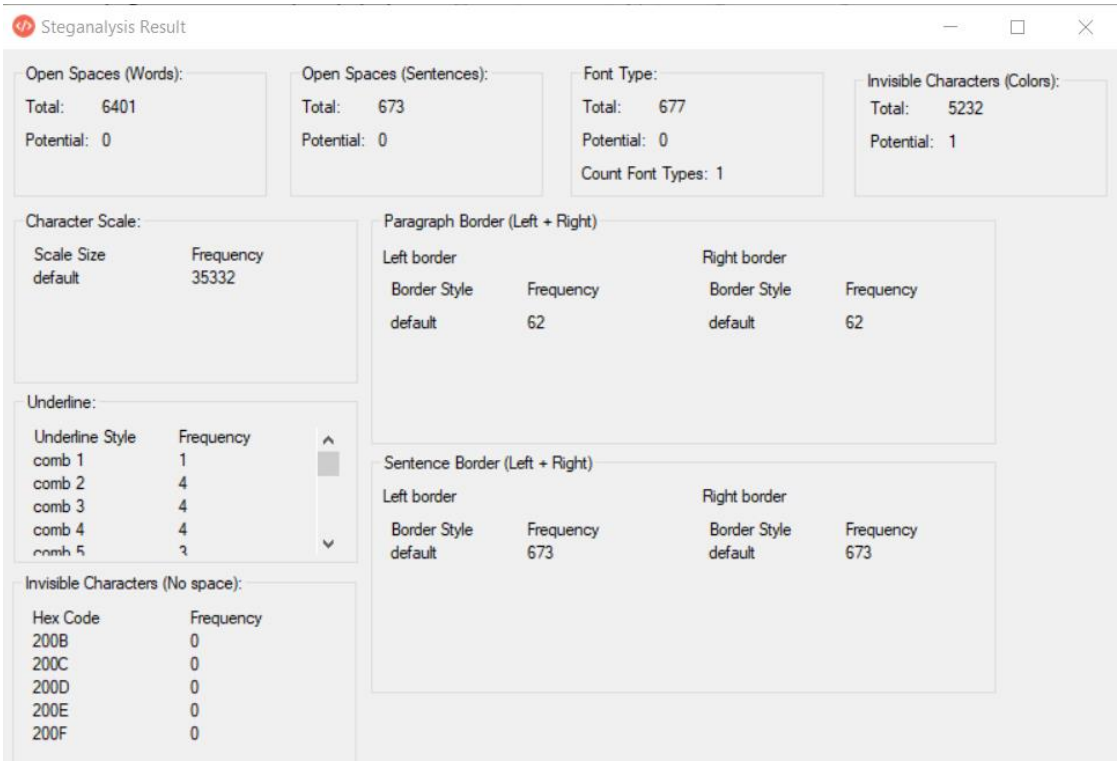
Во продолжение се дадени неколку примери за добиените резултати од извршени детални стеганализи врз *Microsoft Word* документ од 10 страници. Разгледани се резултатите кога во документот е вгнездена скриена порака од 50 знаци, користејќи ги сите четири предложени стеганографски методи, како и анализата на истиот тој документ кога во него нема скриена порака.



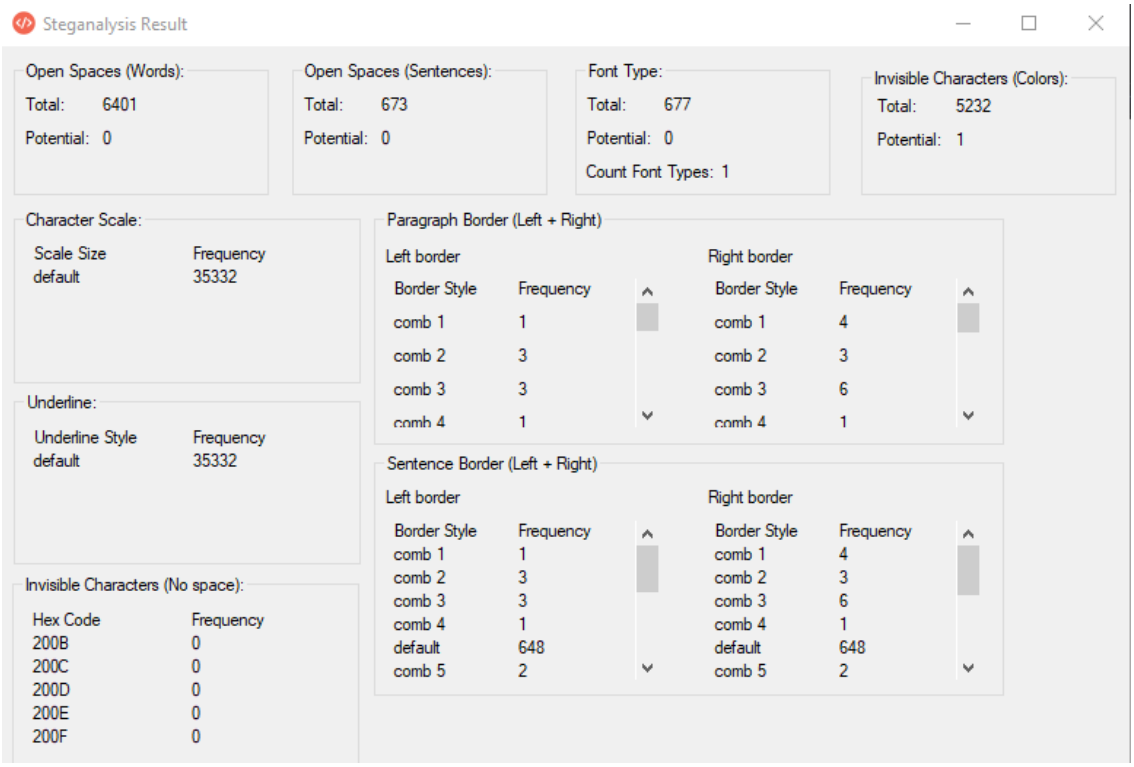
Слика 40: Детална стеганализа врз документ во кој нема скриена порака



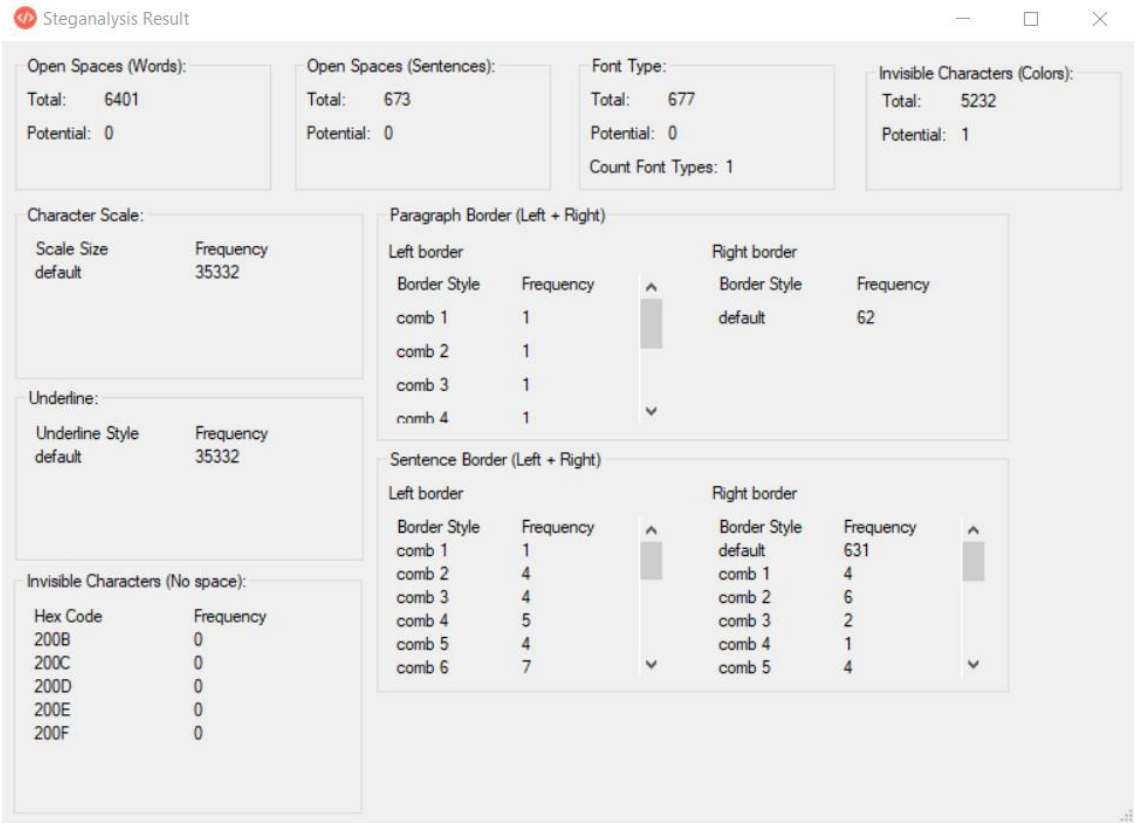
Слика 41: Детална стеганализа врз документ на кој е применет стеганографски метод за скалирање на знаците, со цел вгнездување на скриена порака од 50 карактери



Слика 42: Детална стеганализа врз документ на кој е применет стеганографски метод за подвлекување на знаците, со цел вгнездување на скриена порака од 50 карактери



Слика 43: Детална стеганализа врз документ на кој е применет стеганографски метод за манипулација на границите на параграфите, со цел вгнездување на скриена порака од 50 карактери



Слика 44: Детална стеганализа врз документ на кој е применет стеганографски метод за манипулација на границите на реченициите, со цел вгнездување на скриена порака од 50 карактери

4.1..3 Експерименти

За извршувањето на експериментите е користена компјутерска конфигурација со процесор *Intel(R) Core(TM) i5-3230M CPU @ 2.60GHz 2.60GHz*, *8GB RAM* меморија и *64-bit Windows 10 Home* оперативен систем.

Табела 27: Карактеристики на документите користени во експериментите со методите за стеганографија

	Документ 1	**Документ 2**	**Документ 3**
Страници	1	11	110
Зборови	340	2381	30907
Знаци	2252	15493	190833
Параграфи	13	82	802
Линии	42	328	3445
Реченици	21	134	2028
Оригинална големина (**B**)	31122	923090	4589312

MSWordST: кога станува збор за алатката со која се имплементирани стеганографските методи, направени се експерименти врз три документи: мал, среден и голем, чии што карактеристики се дадени во *Табела 27*.

Во секој од овие документи, се врши вгнездување на скриени пораки со должини од 10, 50, 100, 500, 1000 и 5000 знаци (доколку капацитетот дозволува), со користење на четирите предложени стеганографски методи. Целта е да се анализира големината на документите (пред и после вгнездувањето), на начин што во табелите во продолжение се дадени резултатите за сите три документи. По користењето на четирите методи, се дадени новите големини на документите и процентот за тоа колку големината на новиот документ се зголемила т.е. колку процентуално скриените својства влијаеле на големината на документот.

Скратениците на имињата на методите во табелите се однесуваат на:

- Метод 1 – скалирање на знаците;
- Метод 2 – подвлекување на знаците;
- Метод 3 – манипулација со границите на параграфите;
- Метод 4 – манипулација со границите на реченициите;
- Документ 1 – мал документ;
- Документ 1 – среден документ;
- Документ 1 – голем документ;

Табела 28: Експериментални резултати за стеганографските методи врз Документ 1, со оригинална големина 31122B

	10 карактери		50 карактери		100 карактери		500 карактери		1000 карактери		5000 карактери	
	Големина	%	Големина	%	Големина	%	Големина	%	Големина	%	Големина	%
Метод 1	31448	1.01047	32347	1.03936	33390	1.04074	/	/	/	/	/	/
Метод 2	31249	1.00408	31530	1.01310	31986	1.02776	34482	1.10796	37517	1.20548	/	/
Метод 3	31295	1.00555	/	/	/	/	/	/	/	/	/	/
Метод 4	31356	1.00751	/	/	/	/	/	/	/	/	/	/

Табела 29: Експериментални резултати за стеганографските методи врз Документ 2, со оригинална големина 923090B

	10 карактери		50 карактери		100 карактери		500 карактери		1000 карактери		5000 карактери	
	Големина	%	Големина	%	Големина	%	Големина	%	Големина	%	Големина	%
Метод 1	923609	1.00056	924750	1.00179	925472	1.00258	934834	1.01272	946697	1.02557	/	/
Метод 2	924243	1.00124	924605	1.00164	925180	1.00226	926341	1.00352	928582	1.00624	953474	1.03291
Метод 3	923455	1.00039	924398	1.00141	924547	1.00157	/	/	/	/	/	/
Метод 4	923587	1.00053	924013	1.00099	925290	1.00238	/	/	/	/	/	/

Табела 30: Експериментални резултати за стеганографските методи врз Документ 3, со оригинална големина 4589312B

	10 карактери		50 карактери		100 карактери		500 карактери		1000 карактери		5000 карактери	
	Големина	%	Големина	%	Големина	%	Големина	%	Големина	%	Големина	%
Метод 1	4589321	1.00000	4589363	1.00001	4591027	1.00037	4595001	1.00123	4605370	1.00349	4682285	1.02025
Метод 2	4589313	1.00000	4589356	1.00000	4589574	1.00005	4592093	1.00060	4595782	1.00140	4608077	1.00408
Метод 3	4589512	1.00004	4589567	1.00005	4589597	1.00006	4591231	1.00041	4593443	1.00090	/	/
Метод 4	4589376	1.00001	4589396	1.00011	4591778	1.00010	4595859	1.00142	4603958	1.00319	/	/

Од резултатите во табелите се гледа дека сите четири методи имаат многу мало влијание врз големината на документите т.е. помалку од 1.206% за малиот документ, помалку од 1.033% за средниот документ и помалку од 1.021% за големиот документ. Исто така може да се забележи дека методот со кој се врши подвлекување на знаците има најмало влијание на големината малите и на големите документи, додека пак методот со манипулација на границите на параграфите има најмало влијание кај документите со средна големина.

MSWordSST: кога станува збор за алатката со која се имплементирани методите за стеганализата, направени се експерименти врз два документа: мал и среден, на кои се применуваат четирите предложени стеганографски методи. Документите се пуштени низ процесот на генералната стеганализа од имплементираната алатка, со што во *Табела 31* се дадени резултатите за времето кое е потребно истите да бидат анализирани пред и по вгнездувањето на скриени пораки.

Во сите случаи, времето потребно да се изврши генерална стеганализа е поголемо кога документот содржи скриена порака. Кај малиот документ, генералната стеганализа се извршува најбавно во случај кога е применет стеганографскиот методот за скалирање на карактерите, додека пак кај документот со средна големина, генералната стеганализа се извршува најбавно во случај која е применет стеганографскиот метод за манипулирање со границите на речениците.

Табела 31: Експериментални резултати за времето потребно да се изврши генерална стеганализа (во секунди) пред и по вгнездување на скриени пораки

	Документ со 1 страна	Документ со 10 страни
Документ без скриена порака	60.322	1002.993
Скриена порака од 5 знака, со користење на методот за скалирање на знаци	62.333	1058.471
Скриена порака од 50 знаци, со користење на методот за скалирање на знаци	64.990	1061.845
Скриена порака од 5 знака, со користење на методот за подвлекување на знаци	61.209	1114.241
Скриена порака од 50 знаци, со користење на методот за подвлекување на знаци	61.779	1243.126
Скриена порака од 5 знака, со користење на методот за манипулација со границите на параграфите	61.850	1033.993
Скриена порака од 50 знаци, со користење на методот за манипулација со границите на параграфите на методот за подвлекување на знаци	/	1036.702
Скриена порака од 5 знака, со користење на методот за манипулација со границите на реченициите	61.389	1152.630
Скриена порака од 50 знаци, со користење на методот за манипулација со границите на параграфите на методот за подвлекување на реченициите	62.929	1247.927

5 Заклучок

Стеганографијата и стеганализата се две меѓусебно поврзани науки кои овозможуваат скриена комуникација помеѓу две страни, со користење носачи кои навидум изгледаат сосема обично. Како носачи за пренос на скриената порака може да се јават различни формати како слики, аудио, видео, мрежни протоколи, текст и слично.

Во рамките на овој труд, од аспект на стеганографијата, добиени се следните резултати:

- нов метод за стеганографија, со кој вгнездувањето на скриена порака се врши преку скалирање на знаците во документот;
- нов метод за стеганографија, со кој вгнездувањето на скриена порака се врши преку подвлекување на знаците во документот и манипулација на својствата на линијата за подвлекување;
- нов метод за стеганографија, со кој вгнездувањето на скриена порака се врши преку додавање на границите на параграфите во документот и манипулација со самите својства на границите;
- нов метод за стеганографија, со кој вгнездувањето на скриена порака се врши преку додавање на границите на реченициите во документот и манипулација со самите својства на границите;
- имплементација на предложените методи во форма на алатка (креирана со *Visual C#*) преку која може да се врши вгнездување на скриената порака во даден документ, читање на скриена порака од даден документ и отстранување на скриена порака во даден документ;
- вршење експерименти со документи коишто се користени во креираната алатка за стеганографија, при што е анализирано како вгнездувањето на скриена порака влијае врз големината на документот. Анализата е извршена за сите предложени методи, врз документи со една, единаесет и сто и десет страници;

При предлагањето на секој од методите, за нив се дадени повеќе алтернативи, со цел зголемување на нивните капацитети. Како идна работа во имплементацијата на алатката за стеганографија, се предлага зголемување на капацитетот за пренос, преку имплементација на предложените алтернативи.

Во рамките на овој труд, од аспект на стеганализата, добиени се следните резултати:

- нови методи за стеганализа на отворените методи;
- нов метод за стеганализа на методи што вршат манипулација на невидливите знаци;
- нов метод за стеганализа на методот базиран на типови на фонтови;
- нов метод за стеганализа на методот со скалирање на знаци;
- нов метод за стеганализа на методот со подвлекување на знаците;
- нов метод за стеганализа на методот што врши манипулација со границите на параграфите;
- нов метод за стеганализа на методот што врши манипулација со границите на реченициите;
- имплементација на предложените методи во форма на алатка (креирана со *Visual C#*) преку која може да се врши стеганализа на даден документ и приказ на одредени статистики кои се директно поврзани со својствата коишто стеганографските методи ги менуваат. Крајната цел е врз основа на резултатите, корисникот да може да донесе заклучок за тоа дали во документот постои скриена порака или не;
- вршење експерименти со алатката за стеганализа, врз документи коишто не содржат скриена порака и врз документи коишто содржат скриена порака вгнездена со креираната алатка за стеганографија, при што е анализирано времето потребно за алатката да помине низ документот и да ги прикаже крајните резултати. Анализата е извршена за документи со една и со десет страници, во случај кога тие не содржат скриена порака и во случај кога врз нив се применети четирите предложени стеганографски методи.

Како идна работа во имплементацијата на алатката за стеганализта, се предлага подобрување на имплементацијата, со цел намалување на потребното време за вршење на анализата, кое што во моментот е значително големо кога станува збор за големи документи. Една од алтернативите за постигнување на подобри перформанси е користење на паралелно програмирање во самата алатка. Дополнително се предлага имплементација која што би читала множество од документи, при што стеганализата би се извршила на сите истовремено. Во моментот алатката поддржува читање на еден документ и

негово анализирање, па предложената алтернатива е алатката да поддржува читање на една локација (фолдер) и вршење анализа на сите *MS Word* документи коишто се на таа локација.

Користена литература

[1] J. Trithemius: “Steganographia”, 1499.

[2] M.J. Atallah, V. Raskin, M. Crogan, C. Hempelmann, F. Kerschbaum, D. Mohamed, S. Naik: “Natural language watermarking: design, analysis, and a proof-ofconcept implementation,” Proceedings of the 4th International Workshop on Information Hiding, LNCS vol. 2137, 2001, pp. 185–200.

[3] M.J. Atallah, V. Raskin, C.P. Hempelmann, M. Karahan, R. Sion, U. Topkara, K.E. Triezenberg: “Natural Language Watermarking and Tamperproofing,” Fifth International Workshop on Information Hiding, LNCS vol. 2578, 2003, pp. 196–212.

[4] M. Topkara, C.M., Taskiran, E.J. Delp: “Natural language watermarking,” Proceedings of the SPIE International Conference on Security, Steganography and Watermarking of Multimedia Contents, 2005.

[5] A. Gordon: “Official (ISC)2 Guide to the CISSP CBK”, Fourth Edition, (ISC)2 Press. p. 349. ISBN 1939572061, 2015.

[6] R. Stuart, K. Frederick: “Honor Bound: American Prisoners of War in Southeast Asia, 1961–1973”, Naval Institute Press, ISBN 1-59114-738-7, 2007.

[7] K. Bennett: “Linguistic steganography: Survey, analysis, and robustness concerns for hiding information in text,” cERIAS Tech Report 2004-13, 2004.

[8] S.H. Low, N.F. Maxemchuk, J.T. Brassil, L. O‘Gorman: “Document marking and identification using both line and word shifting,” Proceedings of the 14th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM ‘95), pp. 853–860, 1995.

[9] K. Winstein: “Lexical steganography through adaptive modulation of the word choice hash,”, 1998.

[10] Z. Yang, X. Guo, Z. Chen, Y. Huang, Y. Zhang: “RNN-stega: Linguistic steganography based on recurrent neural networks,” IEEE Trans. Inf. Forensics Secur., vol. 14(5), pp. 1280–1295, 2019.

[11] Y. Luo, Y. Huang, F. Li, C. Chang: “Text steganography based on ci-poetry generation using Markov chain model,” KSII Trans. Internet Inf. Syst., vol. 10(9), pp. 4568–4584, Sep. 2016.

[12] J.T. Brassil, S. Low, N.F. Maxemchuk, and L. O‘Gorman, “Electronic Marking and Identification Techniques to Discourage Document Copying,” IEEE Journal on Selected Areas in Communications, vol. 13, no. 8, October 1995, pp. 1495–1504.

[13] J.T. Brassil, S. Low, and N.F. Maxemchuk, “Copyright protection for the electronic distribution of text documents,” in IEEE, vol. 87, no. 7, July 1999, pp. 1181–1196.
[14] Y. Kim, K. Moon, and I. Oh, “A Text Watermarking Algorithm based on Word Classification and Interword Space Statistics,” in Seventh International Conference on Document Analysis and Recognition(ICDAR‘03), 2003, pp. 775–779.
[15] M. Shirali-Shahreza and S. Shirali-Shahreza: “A New Approach to Persian/Arabic Text Steganography,” Proceedings of the 5th IEEE/ACIS international Conference on Computer and Information Science and 1st IEEE/ACIS, pp. 310–315, 2006.
[16] W. Bender, D. Gruhl, N. Morimoto, and A. Lu, “Techniques for data hiding,” IBM Systems Journal, vol. 35, 1996, pp. 313–336.
[17] A.M. Alattar, O.M. Alattar, “Watermarking electronic text documents containing justified paragraphs and irregular line spacing,” SPIE – Volume 5306, Security, Steganography, and Watermarking of Multimedia Contents VI, June 2004, pp. 685–695.
[18] P.V.K. Borges, J. Mayer, “Document Watermarking via Character Luminance Modulation,” IEEE International Conference of Acoustics, Speech and Signal Processing (ICASSP 2006), vol. 2, 2006.
[19] R. Villán, S. Voloshynovskiy, O. Koval, J. Vila, E. Topak, F. Deguillaume, Y. Rytsar, and T. Pun, “Text Data-Hiding for Digital and Printed Documents: Theoretical and Practical Considerations,” Proceedings of SPIE-IS&T, Electronic Imaging 2007, Security, Steganography, and Watermarking of Multimedia Contents IX, San Jose, USA, 2007.
[20] A.K. Bhattacharjya and H. Ancin, “Data embedding in text for a copier system,” Proceedings of IEEE International Conference on Image Processing - ICIP 99, vol. 2, 1999.
[21] M. Khairullah: “A Novel Text Steganography System Using Font Color of the Invisible Characters in Microsoft Word Documents,” Second International Conference on Computer and Electrical Engineering, pp. 482-484, 2009.
[22] R. Hirsch, Exploring Colour Photography: A Complete Guide. Laurence King Publishing. ISBN 1-85669-420-8, 2004.
[23] A. Odeh, K. Elleithy, M. Faezipour: “Steganography in Text by Using MS Word Symbols,” in Proceedings of the 2014 Zone 1 Conference of the American Society for Engineering Education, 2014.

[24] A. Odeh, K. Elleithy: “Steganography in Text by Merge ZWC and Space Character,” 28th International Conference on Computers and Their Applications, CATA-2013, 2013.
[25] M. Aman, A. Khan, B. Ahmad, S. Kouser: “A hybrid text steganography approach utilizing Unicode space characters and zero-width character,” Int. J. Inf. Technol. Secur. 9, pp. 85–100, 2017.
[26] J. Korpela, IT and communication: Characters and encodings, 2002.
[27] The Unicode® Standard, ISBN 978-1-936213-13-9, 2016.
[28] W. Bhaya, A.M. Rahma, D. Al-Nasrawi: “Text Steganography based on Font Type in MS-Word Documents,” Journal of Computer Science, vol. 9(7), pp. 898–904, 2013.
[29] A.M. Rahma, W. Bhaya, D. Al-Nasrawi (2013): “Text Steganography Based on Unicode of Characters in Multilingual”, International Journal of Engineering Research and Applications (IJERA), vol. 3, no. 4 (1153–1165).
[30] D. DeBlasio, B. Mundt: “An Enhanced Steganographic Method for Data Hiding in Microsoft Word Documents by a Change Tracking Technique”, University of Central Florida.
[31] Y.B. Sadia: “Huffman coding”, Virtual University of Pakistan, 2015.
[32] I. Banerjee, S. Bhattacharyya, G. Sanyal: “Novel Text Steganography through Special Code Generation”, International Conference on Systemics, Cybernetics and Informatics (298-303).
[33] S. Bhattacharyya, I. Banerjee, and G. Sanyal, “A novel approach of secure text based steganography model using word mapping method,” Int. Journal of Computer and Information Engineering, vol.4, pp. 96-103, 2010.
[34] H. Singh, P.K. Singh, K. Saroha (2009): “A Survey on Text Based Steganography”, Proceedings of the 3rd National Conference; INDIACom-2009.
[35] A. Desoky, (2012) ‘Jokestega: automatic joke generation-based steganography methodology’, Int. J. Security and Networks, Vol. 7, No. 3, pp.148–160.
[36] K. Binsted (1996) Machine Humour: An Implemented Model of Puns, PhD Thesis, University Of Edinburgh, Edinburgh, Scotland.
[37] K. Binsted, H. Pain, G. Ritchie, (1997) ‘Children’s evaluation of computer-generated punning riddles’, Pragmatics and Cognition, Vol. 5, No. 2, pp.305–354.

[38] R. Black, A. Waller, G. Ritchie, H. Pain, R. Manurung, (2007) 'Evaluation of joke-creation software with children with complex communication needs', Communication Matters, Vol. 21, No. 1, pp.23–28.
[39] A. Desoky, (2009) 'Listega: list-based steganography methodology', Int. J. Inf. Secur, 8:247–261, pp.247–261.
[40] J. Dénes, A.D. Keedwell: Latin squares and their applications, New York-London: Academic Press. p. 547, 1974.
[41] Z.L. Chen, L.S. Huang, Z.Z. Yu, W. Yang, L.J. Li, X.L. Zheng, X.X. Zhao (2008): "Linguistic steganography detection using statistical characteristics of correlations between words", The 11th International Workshop on Information Hiding, Darmstadt, Germany (224–235).
[42] H. Yang, X. Cao (2010): "Linguistic Steganalysis Based on Meta Features and Immune Mechanism", Chinese Journal of Electronics, Vol.19, No.4, Oct (661–666).
[43] T.J. Jacob, H.G. Gregg, L.C. Roger, G.B. Jr, B. Lamont (2013). Steganography Detection Using a Computational Immune System: A Work in Progress [J]. International Journal of Digital Evidence, 4(1).
[44] Z. Yang, N. Wei, J. Sheng, Y. Huang, Y.-J. Zhang (2018): "TS-CNN: Text Steganalysis from Semantic Space Based on Convolutional Neural Network".
[45] L. Xiang, X. Sun, G. Luo, C. Gan (2007): "Research on Steganalysis for Text Steganography Based on Font Format", Third International Symposium on Information Assurance and Security (490 – 495).
[46] L. Li, L. Huang, X. Zhao, W. Yang, Z. Chen (2008): "A Statistical Attack on a Kind of Word-Shift Text-Steganography", National High Performance Computing Center of Hefei (1503-1507).
[47] E.W. Weisstein: "The CRC Encyclopedia of Mathematics", Third Edition, ISBN 9781420072211, 2009.
[48] I. Stojanov, A. Mileva, I. Stojanovic, "A New Property Coding in Text Steganography of Microsoft Word Documents", Securware 2014: The Eighth International Conference on Emerging Security Information, Systems and Technologies, pp.25 - 30 (2014).
[49] C.M. Taskiran, M. Topkara, E.J. Delp, "Attacks on lexical natural language steganography systems", Proceedings of SPIE - The International Society for Optical Engineering 6072:607209–607209–9 (2006).

[50] X. Zuo, H. Hu, W. Zhang, N. Yu: “Text Semantic Steganalysis Based on Word Embedding”, Sun X., Pan Z., Bertino E. (eds) Cloud Computing and Security (ICCCS 2018), LNCS vol.11066, Springer, Cham (2019).
[51] J. Wen, X. Zhou, P. Zhong, Y. Xue, “Convolutional Neural Network Based Text Steganalysis”, IEEE Signal Processing Letters, Vol. 26(3) (2019).
[52] Z. Yang, N. Wei, J. Sheng, Y. Huang, Y.J. Zhang: “TS-CNN: Text Steganalysis from Semantic Space Based on Convolutional Neural Network”, (2018).
[53] Z. Yang, K. Wang Li, Y. Huang, Y.J. Zhang, “TS-RNN: Text Steganalysis Based on Recurrent Neural Networks”, IEEE Signal Processing Letters, Vol. 26(12) (2019).
[54] Z. Yang, Y. Huang, Y.J. Zhang: “TS-CSW: text steganalysis and hidden capacity estimation based on convolutional sliding windows. Multimedia Tools and Applications”, (2020).
[55] Z. Yang, Y. Huang, Y.J. Zhang: “A Fast and Efficient Text Steganalysis Method”, IEEE Signal Processing Letters, Vol. 26(4) (2019).
[56] I. Stojanov, A. Mileva, D. Stojanov, N. Stojkovik: “MSWordSST - A New Steganalytical Tool for Microsoft Word Documents,” 12th ICT Innovations Conference, 2020.

Иван Стојанов

КРИЕЊЕ НА ПОДАТОЦИ ВО ЕЛЕКТРОНСКИ ДОКУМЕНТИ

Универзитет „Гоце Делчев“ - Штип