Исследование методов текстовой стеганографии

Текстовая стеганография может варьироваться от изменения форматирования существующего текста до изменения слов внутри текста и создания нового текста. Для создания понятных текстов используются случайные последовательности символов или контекстно-свободные грамматики.

Из-за отсутствия избыточной информации, содержащейся в файлах изображений, аудио или видео, текстовая стеганография считается самой сложной. Структура текстовых документов идентична тому, что мы видим, тогда как структура других видов документов, таких как изображения, отличается от того, что мы видим. В результате мы можем скрыть информацию в таких документах, изменив структуру документа, не влияя на вывод.

Изображение или аудиофайл можно изменить способами, которые невозможно обнаружить; однако случайный читатель может пометить текстовый файл дополнительной буквой или знаком препинания. Текстовые файлы требуют меньше памяти для хранения, и они быстрее и проще в передаче, чем другие формы стеганографических технологий.

Текстовую стеганографию можно разделить на три категории: лингвистические подходы, случайная и статистическая генерация на основе формата. На рисунке 3 показан механизм текстовой стеганографии.

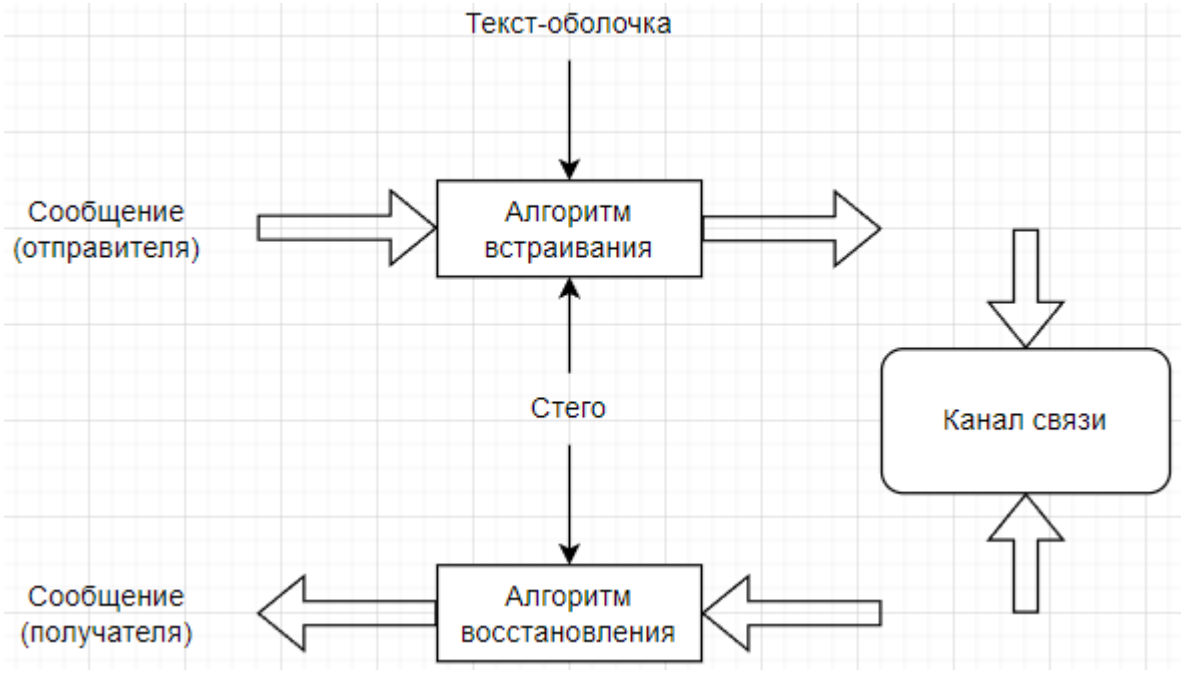

Рисунок 1 – Механизм текстовой стеганографии

Правописание слов

Мы используем этот процесс для сокрытия информации на английском языке. В этих процессах написание слов в американском (США) языке объясняется иначе, чем в британском (Великобритания). Приведем пример: у Organize есть другие варианты написания в Великобритании (Organise) и США (Organize). Этот процесс подходит для области, где редко используются как американские, так и британские термины. В этом процессе, если кто-то не знает метода обоих написаний, но легко обнаруживает, то наши данные скрыты.

Есть небольшая разница между написанием слов в американском и британском языках, например, в американском языке (Color) и в британском языке (Colour), небольшие изменения только в «u», которые легко обнаружить. Метод New Synonym Text является более конфиденциальным процессом, чем этот метод, поскольку в методе New Synonym мы используем различные слова, такие как (Soccer) в американском языке и (Football) в британском языке. Очевидно, что те, кто использует эти два слова как синонимы, не смогут легко понять.

Семантический метод

Этот процесс такой же, как и процесс написания слов. Хотя небольшое отличие заключается в том, что в этой технике мы можем использовать слова, которые являются синонимами слов, поэтому используются специальные слова, которые скрывают информацию в тексте. Это также защищенные данные (OCR) в случае применения или использования программы идентификации. У нас есть много преимуществ. Поэтому нужна самая надежная техника, потому что если у кого-то есть много приказов или команд на его языке, он может легко обнаружить скрытые данные.

Метод смещения строк

В этом методе строки текста перемещаются на разную величину (например, каждая строка перемещается на 1/300 дюйма вверх и вниз), и путем обнаружения идентичной формы текста информация скрывается.

Это устройство для измерения расстояния и обязательная модификация будут инициированы для устранения скрытой информации. Кроме того, если используется программа распознавания символов (OCR), напечатанные данные повреждаются. Этот метод полезен для печатного текста, поскольку OCR не используется в печатных тестах.

Метод сдвига слов

В этом методе конфиденциальное сообщение скрывается путем преобразования слов по горизонтали, так что справа или справа не будет отображаться небольшой 0 и 1 соответственно, а в тексте данные скрываются путем изменения расстояния между словами. Эта практика подходит для текстов, где интервал между словами не одинаков. Этот процесс можно недооценить, так как обычно интервал между словами изменяется для заполнения строки. Но если кто-то знает об алгоритмах, связанных с методом сдвига слов, он легко получит скрытые данные.

Синтаксический процесс

В этом процессе, изменяя некоторые знаки препинания, такие как запятая (,), точка (.), точка с запятой (;), кавычки (“) в соответствующем положении, любой может скрыть данные в текстовом файле. Требования этого процесса — идентификация соответствующих мест или вкладов. Преимущество этого метода в том, что для защиты данных практически не требуется информации. Например: в любом стихотворении или абзаце мы определяем точку (.) как 0, а запятую (,) как 1 для защиты данных в нем и отправки их пользователям.

Новый метод синонимического текста

В этом процессе некоторые слова с их синонимами используются для защиты секретного сообщения в текстовом файле. В методе правописания слов написание меняется очень мало, но в новой синонимической технике для одного и того же числа используются различные типы слов. Некоторые слова в английском языке имеют разные термины в Соединенных Штатах (US) и Соединенном Королевстве (UK). Например, «Movie» имеет другой термин в Великобритании как (Flim) и в США как (Movie).

Этот процесс более полезен, чем метод правописания слова, потому что в этой технике используются различные типы слов, которые нелегко понять. В технике правописания используются различные типы написания слов, например, в США (faculty) и в Великобритании (staff).

Недостатком этого метода является то, что он занимает немного времени, поскольку нам приходится искать синонимы слов и заменять их, пока не получим соответствующие результаты [4]. На рис. 3 показано несколько приемов сокрытия сообщения.

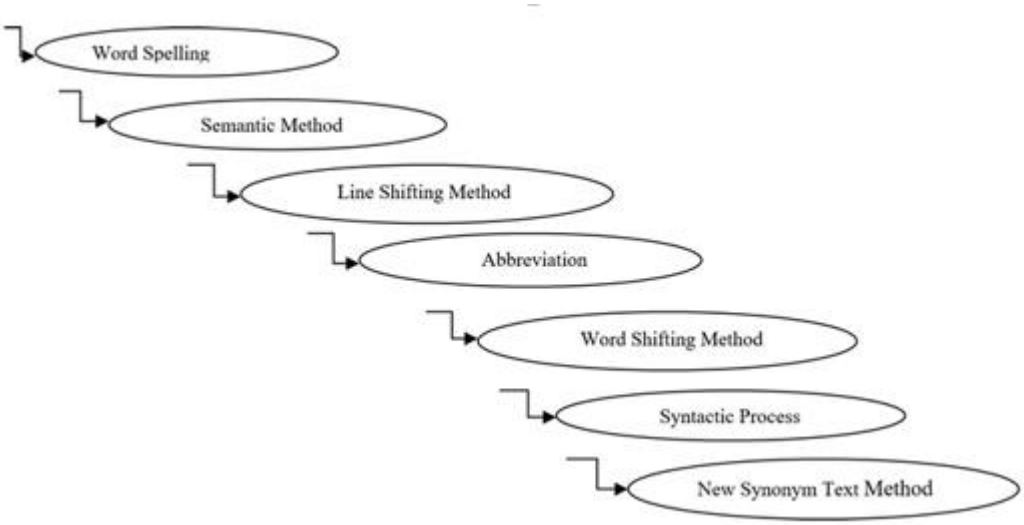

Рисунок 2

Механизм сокрытия текста

Поскольку читать может каждый, шифрование текста в беспристрастных предложениях вряд ли будет эффективным. В предыдущем предложении, если вы возьмете первую букву каждого слова, вы увидите, что это не невозможно и очень просто.

Существует много способов защитить информацию в текстовом файле. Алгоритм первой буквы, используемый здесь, очень ненадежен, потому что знание используемой системы автоматически раскрывает вам секрет. Недостатком является то, что существует много распространенных методов сохранения секретов в простом тексте.

Многие методы включают такие правила, как изменение порядка текста, использование каждого девятого символа или изменение количества пробелов после строк или между словами. Успешно использовался последний метод, и даже после печати и копирования текста на бумаге десять раз зашифрованное сообщение можно восстановить.

Другой эффективный способ шифрования текста — использовать общедоступный источник обложки, книгу или газету и использовать код, который содержит, например, номер строки, номер страницы, букву и номер

включения. Таким образом, никакая информация, скрытая внутри сервера, не приведет к скрытому сообщению.

Его обнаружение полностью зависит от получения знаний о секретном ключе. Сосредоточение на методе внедрения, используемом для сокрытия конфиденциальной информации в тексте обложки. Существует три основные категории текстовой стеганографии, которые являются методами на основе формата, случайной и статистической генерацией и лингвистическими методами.

Методы на основе формата

Физическое форматирование текста способами на основе формата используется как место для сокрытия данных. Методы на основе формата обычно редактируют существующий текст, чтобы скрыть стеганографический текст. Метод стеганографии текста на основе формата также известен как метод открытого пространства.

Взаимодействующие пробелы или невидимые символы, преднамеренное прописывание всего распределения текста и изменение размера шрифтов — вот некоторые из многих методов форматирования, используемых в текстовой стеганографии.

Некоторые из этих процессов, такие как преднамеренные орфографические ошибки и интервалы, иногда могут обмануть читателей-людей, которые игнорируют неправильные толкования, но часто могут быть легко обнаружены машиной.

Случайные и статистические методы

Чтобы избежать сравнения с известным простым текстом, стенографисты всегда поддерживают подготовку собственных текстов сокрытия. Хотя это часто решает проблему атаки на известное ядро, особенности подготовленного текста все еще могут вызывать подозрения, что текст является незаконным.

Такая генерация обычно пытается имитировать некоторые особенности общего текста, приближаясь к некоторым статистическим разделениям, обнаруженным в исходном тексте.

Текст может скрывать информацию от стенографии до точки зрения, которая случайным образом показывает ряд символов. Конечно, эта настройка далека от случайной как для отправителя, так и для получателя сообщения, но она должна быть случайной для всех, кто перехватывает сообщение.

Однако она не только должна быть известна случайным образом, но поскольку нас также беспокоит тот факт, что это стенографическая фраза, она не выглядит подозрительной. Случайные наборы символов, которые все попадают в один и тот же набор символов, но не имеют четкого значения, могут действительно вызвать предупреждение.

Кодирование признаков

Кодирование признаков имеет дело с изменением признаков текста таким образом, что значимое сообщение скрывается для создания текста-обложки. Такие признаки, как высота текста, цвет текста, шрифт текста, являются некоторыми из способов, которые используются. Большой объем информации скрывается с помощью кодирования признаков.

Когда признаки текста изменяются, то только отправитель и предполагаемый получатель обнаруживают скрытое сообщение. Третья сторона не привлекает внимание к чему-то скрытому внутри текста. Методы OCR и перепечатка ответственны за изменение признаков текста и повреждение сообщения.

Текстовая стеганография в языке разметки

Язык разметки — это современная система для аннотирования документа таким образом, чтобы он был синтаксически отличим от текста. Языки разметки, нечувствительные к регистру, можно легко использовать для создания скрытых смыслов. Для этой цели используется HTML. Теги html не поддерживают чувствительность к регистру. Теги, такие как
 и
,

дают одинаковый вывод, независимо от регистра, в котором они написаны. Другие теги, такие как <b>,<u> и т. д., могут использоваться для того, чтобы было больше возможностей для стеганографии в HTML. XML является чувствительным к регистру языком. Это означает, что теги, написанные в XML, зависят от заглавных букв алфавита. Для целей стеганографии в основном используется HTML. Он увеличивает диапазон, в котором отправитель использует функциональность создания секретного сообщения.

Последовательности слов

Проблема, возникающая в том, что обнаружение текста обложки в обычном тексте иногда не так уж и сложно. Могут быть обнаружены различные нелексические последовательности, и мощность стеганографии уменьшается. Для решения этой проблемы фактические элементы обнаружения могут использоваться для кодирования одного или нескольких бит информации на слово. Сопоставление между лексическими последовательностями и последовательностями битов может потребовать кодовой книги. Это связано с тем, что биты в текстах используются для кодирования лексического сообщения. Этот метод также имеет несколько проблем, заключающихся в том, что и человек, и компьютер могут обнаружить строку слов без семантической структуры. Аномальное поведение может привлечь злоумышленников и разрушить суть стеганографии.

Последовательности символов

В тексте есть несколько символов. Прелесть текстовой стеганографии заключается в том, что символы языков могут быть использованы для создания текста обложки. В методе последовательности символов генерация символов заключается в учете свойств длины слова и частоты букв для создания слов. Это придает внешнему виду те же статистические свойства, что и фактические слова в данном языке.

Лингвистические методы

Фактические, оригинальные элементы словаря могут использоваться для кодирования одного или нескольких битов информации в каждом слове, чтобы решить проблему идентификации небуквальной непрерывности.

Это может включать в себя кодовую книгу карты между лексическими объектами и битовыми конфигурациями или словами (длинами, буквами и т. д.), кодирующими скрытую информацию. Однако в обоих случаях есть проблема. Строка слов, не имеющая семантической диаграммы и понятной семантической связи. И люди, и компьютеры могут знать одно и то же. Этот метод требует правильной идентификации мест, куда могут быть вставлены знаки. Другой метод лингвистической стеганографии — семантический метод. В этом методе используются синонимы слов для некоторых предварительно выбранных. Слова заменяются их синонимами, чтобы скрыть в них информацию. На рисунке 5 показана классификация текстовой стеганографии.

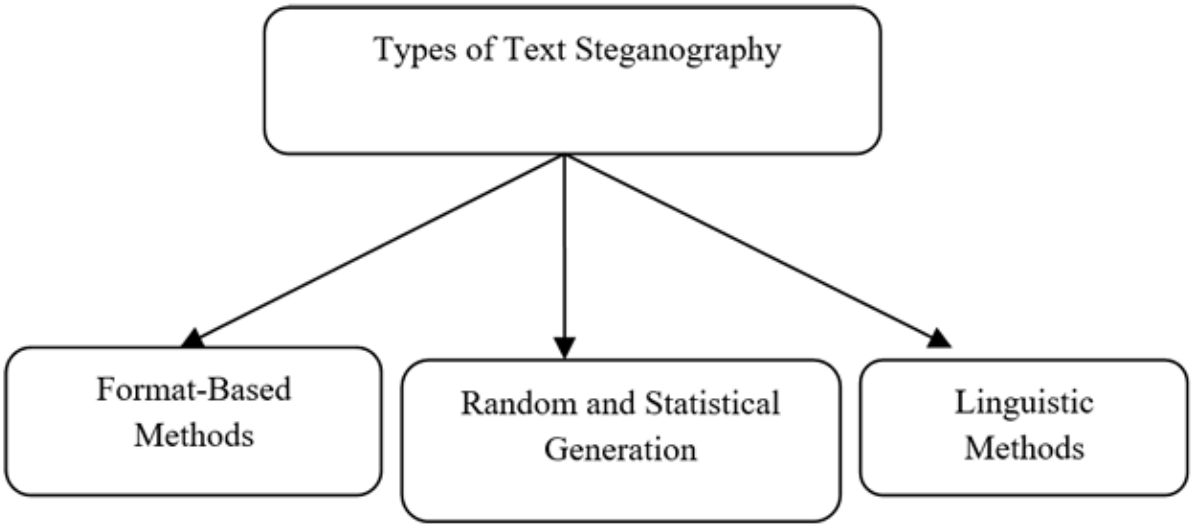

Рисунок 3 – Классификация текстовой стеганографии