В.Й. Залевський, с.н.с
Житомирський військовий інститут ім. С.П. Корольова

ПІДХІД ДО АВТОМАТИЧНОЇ ОБРОБКИ РЕЗУЛЬТАТІВ МОНІТОРИНГУ КОРОТКИХ ТЕКСТОВИХ ПОВІДОМЛЕНЬ В МЕРЕЖІ ІНТЕРНЕТ В ІНТЕРЕСАХ ІНФОРМАЦІЙНОЇ БЕЗПЕКИ ДЕРЖАВИ

Останнім часом в мережі Інтернет значна кількість повідомлень, явищ, процесів описується електронними засобами масової інформації (е-ЗМІ) або обговорюється у соціальних мережах. Ця інформація найчастіше описується короткими текстовими повідомленнями та може породжувати множини неконтрольованих внутрішніх та зовнішніх загроз, які негативно відбиваються на стані інформаційної безпеки держави.

Моніторинг та накопичення масивів таких повідомлень активно використовуються при спостереженні за оточенням, станом об'єктів, суспільною думкою. Сучасний розвиток мережевих засобів масової інформації, а також засобів автоматичного збирання текстової інформації мережі Інтернет дозволяє інформаційним службам отримувати тисячі повідомлень за день. Інформаційні потоки великої інтенсивності і розмірності унеможливлюють ознайомлення аналітика з кожним повідомленням і розумінням його сенсу.

Одним з варіантів розв'язання цієї проблеми є використання засобів автоматичної класифікації текстових повідомлень, одержуваних автоматизованими системами збору інформації з мережі Інтернет. Необхідною умовою ефективної роботи подібних систем є застосування алгоритму, що забезпечує необхідну якість класифікації.

Проблемним питанням автоматизованої та автоматичної класифікації текстів, що надходять з різних джерел інформації приділялася увага багатьох вчених. Ними запропоновано метод автоматичної класифікації коротких текстових повідомлень на основі характеристики тематичної значущості тексту і її модифікації. Показані високі результати класифікації коротких рекламних повідомлень. Однак дані результати були отримані при класифікації тільки за п'ятьма рубриками, що при обробці повідомлень за напрямом інформаційних загроз державі є явно недостатнім. Для вирішення такої задачі виникає необхідність у використанні ієрархічного рубрикатора, з декількома рівнями вкладень.

У доповіді розглядається підхід до створення автоматичної системи класифікації текстових повідомлень на основі модифікованого методу Байеса, що дозволяє в цілому зберегти переваги базового методу, за умови підвищення якості класифікації. Доцільність застосування обраного методу полягає й у тому, що отримані короткі повідомлення представляють собою незалежні величини, подальша модифікація яких не потребує багато зусиль. Відмінність від основного полягатиме у тому, що слова рубрики попередньо будуть піддаватися процедурі додаткової обробки, спрямованої на визначенні значущих (ключових) слів для рубрики з використанням процедур «TF-IDF» та його модифікацій (TF-IDF,TF-RF,TF-ICF).

В якості інформаційної бази для створення даної системи авторами використовувалися структуровані масиви коротких текстових повідомлень з е-ЗМІ та соціальних мереж Вконтакте, Twitter, Facebook.

У запропонованій системі запропоновано точний підхід до морфологічному аналізу, заснований на використанні словників, у яких для кожного слова зазначено правило зміни його форми. В результаті застосування морфологічного аналізу текстових повідомлень формується перелік слів в тексті у початковій формі. Далі до отриманого переліку застосовується процедура видалення «стоп-слів», що являють собою малоінформативні слова і службові частини мови, до того ж не характеризують текстові повідомлення за змістом. Наприклад це можуть бути прийменники, сполучники тощо. Для даної процедури застосовується словник стоп-слів та слів з структурою до трьох символів. Алгоритм обробки повідомлень в загальному вигляді представлено на рисунку 1.

Робота з класифікатором полягатиме у його навчанні по заздалегідь відібраними текстовими повідомленнями та створення векторів текстових повідомлень навчальної вибірки за кожною темою рубрикатора. При цьому ваги термінів мають наступні властивості:

високі значення, якщо термін часто зустрічається в невеликому числі текстових повідомлень, тим самим посилюючи відміну цих повідомлень від інших;

низькі значення, якщо термін рідко зустрічається в якомусь текстовому повідомленні або зустрічається в багатьох повідомленнях, тим самим знижуючи відмінність між ними.

Результатом застосування процедур "TF-IDF, TF-RF, TF-ICF» до векторів текстових повідомлень навчальної вибірки (для кожної теми рубрикатора) є результуючий вектор рубрики з елементами оцінки за тональностями.

Рис. 1. Алгоритм обробки повідомлень

Алгоритм класифікації дозволяє проводити пошук найбільш імовірної рубрики зазначеного повідомлення. Класифікація текстового повідомлення проводиться у два етапи.

На першому етапі здійснюється його первинна обробка. На виході першого етапу повідомлення представляється у вигляді вектора слів в початковій формі.

На другому етапі здійснюється порівняння ознак даного вектора з векторами рубрик навченого рубрикатора. Результати порівняння зберігаються.

Таким чином, формується масив даних, що відображають відповідність повідомлень кожної з рубрик. Далі проводиться його ранжування і переклад значень ступеня відповідності в шкалу від 0 – 1. На виході другого етапу у долях формується масив, що містить відповідність повідомлення кожній рубриці.

Розглянуті процедури обробки коротких текстових повідомлень реалізовані в системі автоматичної класифікації коротких текстових повідомлень «МОНІТОР». Дана система дозволяє провести збір та первинну обробку текстових повідомлень, провести навчання класифікатора до опрацювання повідомлень згідно рубрикатора і в подальшому регулювати розмір результуючого масиву, включаючи в нього інформацію по рубрикам.

Система реалізована з використанням відкритого програмного коду на мові Python. Середовище для зберігання текстових повідомлень організоване на базі СУБД PostgresSQL.

Для аналізу якості класифікації був побудований рубрикатор напрямків загроз інформаційній безпеці держави за 12 напрямками і сформована навчальна і контрольна вибірки коротких текстових повідомлень для їх навчання і оцінки точності класифікації.

Точність класифікації за умові належності до однієї рубрики з дванадцяти складала 81,2 %. До двох найближчих рубрик – 83,1%. До трьох найближчих рубрик – 85,4%.

Апробація розробленої програмної системи групою фахівців аналітичного відділу з обробки первинної інформації показала можливість його застосування у оперативному аналізі коротких повідомлень, що розповсюджуються в е-ЗМІ для виявлення повідомлень які можуть становити загрозу інтересам інформаційної безпеки держави.