УДК 519.767.6

Колесник А.С. (НТУ «ХПІ»)
д.т.н., проф. **Хайрова Н.Ф.** (НТУ «ХПІ»)

DOI: https://doi.org/10.17721/2519-481X/2019/65-05

ВИКОРИСТАННЯ МЕТОДІВ ЯКІСНОГО КОНТЕНТ-АНАЛІЗУ ДЛЯ ДОСЛІДЖЕННЯ ІМОВІРНІСНОГО РОЗПОДІЛУ КРИМІНАЛЬНО ЗНАЧИМОЇ ІНФОРМАЦІЇ НА НОВИННИХ ВЕБ-САЙТАХ

Сьогодні, в століття розвитку інформаційного суспільства, ЗМІ грають роль потужного засобу формування громадської думки і впливу на нього. Зараз усю інформацію можна знайти в текстовому виді в інтернеті, у тому числі і через ресурси соціальних медіа. Для аналізу такого роду даних краще всього застосовувати актуальну на сьогодні інформаційну технологію таку як контент-аналіз, який вивчає документи в їх соціальному контексті і використовується при дослідженні тематичної спрямованості засобів масової інформації. Завдяки розвитку методів контент-аналізу з'явилася можливість автоматичного дослідження змісту текстів, їх ефективності і оцінки впливу на суспільство. У цьому дослідженні проводиться аналіз існуючих підходів, методів і інструментів контент-аналізу, а також обгрунтовується актуальність дослідження можливостей використання широкого спектру лінгвістичних категорій для якісного контент-аналізу. Розглядаються концептуальні можливості використання цього виду аналізу в сучасних лінгвістичних і соціальних дослідженнях. У статті показано використання методів якісного контент-аналізу, підходів machine learning і розробленого словника кримінально забарвлених термінів на трьох мовах, що є одним з основних інструментів для дослідження розподілу кримінально значимої інформації новинних сайтів веб-медіа за географічними, часовими ознаками та категоріями злочинів. В дослідженні також пропонуються базові основи розробки інформаційної технології контент-аналізу новинного веб-простору певних географічних регіонів в часовій залежності за заданою тематикою, а саме по кримінальній картині регіону. В якості експериментального корпусу розглядаються тексти, що були зібрані автоматично, за допомогою розробленої програмного продукту, з новинних сайтів Казахстану, України, Великобританії і США.

Ключові слова: якісний контент-аналіз; новинні веб-сайти; кримінально значима інформація; засоби масової інформації; казахський, український і англійський текстові корпуси, словник кримінально-забарвлених термінів; методи machine learning.

Вступ. Сьогодні одним з найбільш популярних напрямів дослідження великих об'ємів текстових даних став контент-аналіз. І незважаючи на те що він до сих пір має більшу популярність в області соціальних та гуманітарних наук, цей підхід є універсальним, оскільки багато сучасних досліджень спрямовані на практичне і максимально автоматизоване застосування методів контент-аналізу. У наш час контент-аналіз використовується при проведенні соціальних досліджень, опитувань, для виявлення шаблонів та тенденцій змін в громадській думці у хронологічному зрізі. Отримані таким чином дані можна порівнювати в різних проміжках часу і виявляти зміни, що відбуваються в суспільстві, політиці або культурі [1]. Також контент аналіз є одним з найпростіших способів, що часто використовується для визначення авторства. Методи контент-аналізу є актуальними також при вивченні документів в їх соціальному контексті, що дозволяє виявити те, що може вислизнути при традиційному вивченні, статей, публікацій ЗМІ, промов відомих людей і тд.

Метою контент-аналізу є створення кількісного опису смислового і символічного змісту тексту документу, що дозволяє фіксувати об’єктивні ознаки і здійснює їх підрахунок. Контент-аналіз дозволяє "вписати" зміст документу в соціальний контекст, осмислити його одночасно і як прояв, і як оцінку соціального життя. Такий аналіз документу дозволяє виявити, що: (1) існувало до нього і отримало в ньому відображення; (2) існує в ньому (характеристики форми - мова, структура, жанр повідомлення, ритм і тон мови); (3) буде після нього (оцінка різних ефектів дії).

В цілому, контент-аналіз може бути використаний як: основний, паралельний, контрольний (допоміжний) метод. Основний метод використовується при визначенні тематичної спрямованості досліджуваного матеріалу. Паралельний метод застосовується у поєднанні з іншими методами, а контрольний чи допоміжний при класифікації відповідей на анкетні питання.

Важливою відмінністю контент-аналізу від інших подібних методів є те що він чіткий, формалізований і систематизований. Суть його систематизованості та формалізованості полягає в необхідності чіткого визначення категорій аналізу (ключових понять документу), що дозволяє визначати цей метод як досить чіткий та систематизований.

Після визначення системи категорій аналізу необхідно обрати одиницю аналізу тексту, що відповідає цим категоріям. За одиницю аналізу може бути прийнято: слово, речення, тема, ідея, автор, персонаж, соціальна ситуація і інші елементи.

Наступним етапом аналізу є встановлення одиниці рахунку, що являє собою елемент кількісної міри одиниці аналізу, який дозволяє реєструвати частоту (регулярність) появи ознаки категорії аналізу в тексті (число друкарських знаків, сторінок, абзаців, авторських листів). Останній етап це експертний аналіз отриманих результатів дослідження.

Таке представлення контент-аналізу вимагає детального попереднього аналізу мети методу і визначення її залежності від вибраних категорій та їх кількісного або якісного значення.

Огляд літературних джерел з використання контент-аналізу у сучасних лінгвістичних та соціальних дослідженнях. Контент-аналіз – це універсальний метод для роботи з інформацією, що використовується з різними видами джерел даних, включаючи текстові, візуальні і аудіодані [2]. Ця методика дуже гнучка і використовується як емпірично так і теоретично [3]. В цілому контент-аналіз можна розділити на якісний і кількісний.

У простому випадку кількісний контент-аналіз включає підрахунок слів або речень. У складнішому це процес виділення категорій в даних (репрезентативних зразків змісту), їх кодування, яке дозволяє виявити або відобразити ті або інші відмінності в аналізованих даних, а потім оцінка точності цього кодування. Зібрані дані аналізуються для опису шаблонів чи характеристик контенту або для ідентифікації зв'язку між розглянутими властивостями даних [4].

Кількісний підхід добре себе зарекомендував в так званій предиктивній поліцейській діяльності. Вона полягає в застосуванні кількісних аналітичних методів для визначення перспективних цілей запобігання злочинів. При аналізі використовуються вже наявні дані по злочинах (записи допитів, архівні дані, протоколи і тд). У дослідженні [5] показано, що цей підхід дозволяє передбачити з високою ймовірністю місце і час правопорушення, а також виявити людей, що знаходяться в групі ризику.

Слід зазначити, що досить велика кількість досліджень присвячена використанню методів контент-аналізу в задачах криміналістики. Наприклад, кількісні лінгвістичні показники, що використовуються в дослідженні [6] у якому проведено контент-аналіз текстових записів 30 людей, які вчинили самогубство або здійснили терористичні злочини, дозволили виявити тенденції (географічні, наявність групової приналежності), що вказують на злочин.

При виборі між кількісним і якісним аналізом, сучасні дослідники схиляються до якісного, оскільки він використовує індуктивні міркування за допомогою яких категорії виникають завдяки ретельному аналізу і постійному порівнянню [7], яке проводить дослідник [8]. Наприклад, кількісний контент-аналіз успішно використовується для визначення близької тематики в творах авторів, при якому визначається частота появи заданих слів і фраз, а якісний проводить глибший аналіз тексту, що базується на звя’зках між поняттями.

Дослідження, що використовують якісний підхід, фокусуються на характеристиках мови як об'єкті комунікації. У цьому випадку контент-аналіз не зупиняється на простому підрахунку слів, він переходить на рівень інтенсивного вивчення мови, метою якого є класифікація великої кількості категорій, які можуть відображати явне або передбачуване повідомлення.

Можна виділити 3 підходи до якісного аналізу [9]. Перший з них це традиційний, при якому категорії для кодування виводяться з необроблених даних. Він використовується при розробці теорії. Також є спрямований (категорії для кодування беруться з теорії або результатів дослідження, а потім дослідники використовують ці дані і створюють категорії з даних) та підсумковий (на початку здається кількісним оскільки розпочинається з підрахунку слів, а потім розширює аналіз щоб розкрити теми (категорії)) [10].

Сам процес контент-аналізу досить багаторівневий, але незалежно від того застосовується він емпірично або ж теоретично, одним з важливих етапів є етап "створення словника", що використовується для категоризації. Під словником мається на увазі список слів, який асоціюється з тією або іншою категорією. Вони можуть створюватися вручну великою кількістю експертів або ж автоматично за допомогою різних методів machine learning комп'ютерної лінгвістики. Так звані мануальні словники широко застосовуються на маленьких текстах і незначному об'ємі даних. Можливе використання вже існуючих словників, наприклад, добре відомих словників позитивних і негативних слів [11].

В той же час, найпривабливішим способом створення словника є автоматичний. Для його використання потрібен великий набір текстів (наприклад, повідомлення з соцсетей, статті, твіти), які розмічені якимсь чином, або характеризовані по атрибутах людини (стать, вік і т.д). Після чого ці тексти необхідно розбити на слова або словосполучення за допомогою токенізації, а далі застосувати, наприклад, стохастичні методи для визначення слів або фраз, найбільш пов'язаних з результатом.

Ще один спосіб створення словника грунтується на "краудсорсінгу", що дозволяє збирати велику кількість слів без особливих зусиль і витрат за часом. Використовуючи рейтинги можна створювати "зважені словники", в яких кожне слово має відповідну вагу в певній категорії (наприклад, слово "приголомшливий" може бути оцінене як 4.6 за 5-бальною шкалою позитивних емоцій, тоді як слово "хороший" може отримати тільки 3.5). Така оцінка практично сотнями або навіть тисячами людей дозволяє розробити словник, який може охопити навіть слова з категорії загальних і нейтральних слів.

І кількісний і якісний підходи контент-аналізу мають широкий спектр застосування. Наприклад, в дослідженні [12] увага приділяється контент і сентимент аналізу, присвяченому темі вірусу Еболи в стрічці новин і Твіттері, який проводився за допомогою словникового контролю зібраних даних; використання n- граммної техніки моделювання тем LDA; аналізу сутностей і їх мережі, а також використання системи оцінок тем. Результати показали, що освітленість цієї теми в Твіттері більш розмита ніж в новинах. Крім того, в статтях з новин більше уваги приділяється об'єктам пов'язаним з подіями, таким як особа, організація і місце розташування, тоді як Twitter охоплює об'єкти орієнтовані на якийсь час.

Окрім соціальної і кримінальної сфери контент-аналіз активно використовується в еколінгвістиці. У дослідженні [13] був здійснений контент-аналіз публікацій з різних журналів, пов'язаних з еколінгвістикою, що показує різке підвищення уваги до неї.

Інструменти, що використовуються для автоматизованого контент-аналізу

Задача контент-аналізу вирішується у рамках загального підходу (NLP). Обробка природної мови є дуже широкою областю, що дозволяє отримувати нові результати із вже наявних текстових даних. Істотна частина технологій NLP базується на методах глибокого навчання (deep learning) - одній з областей машинного навчання (machine learning). Притому методи машинного навчання добре працюють тільки за наявності розроблених людиною представлень (representations) цих і вхідних ознак, а також оптимізації вагів, що дозволяють зробити фінальний висновок кращим. Тоді як при глибокому навчанні алгоритми автоматично витягають кращі ознаки або представлення з сирих вхідних даних.

Напрям word embedding і text embedding представляють важливі методи NLP при яких слова зіставляються з векторами дійсних чисел. Векторне представлення часто є стартовою точкою для NLP завдань. Метод word embedding дозволяє формалізувати значення слова в документі, виявити семантичну і синтаксичну схожість, а також визначити зв'язок з іншими словами. Він застосовується в завданнях класифікації текстів, отже, є незамінним інструментом для контент-аналізу.

Word2vec є одним з найпопулярніших методів реалізації Word embedding з використанням двошарової нейронної мережі. На вхід поступає текстовий корпус, а на вихід - набір векторів. Багато дослідників вважають, що Word embedding через word2vec може зробити природну мову такою, що прочитується комп'ютером, для виявлення схожості між словами і текстами з можливістю реалізації математичних операцій над векторами. Добре навчений набір векторів слів помістить схожі слова близько один до одного в VSM просторі. Наприклад, слова "помідор", "огірок" і "овочі" можуть об'єднуватися в одному кутку, а жовтий, червоний і синій - в іншому.

Word2vec використовує два основні навчальні алгоритми, один - continuous bag of words (CBOW), інший - skip - gram. Основна відмінність між цими двома методами полягає в тому, що CBOW використовує контекст для прогнозування цільового слова, тоді як skip - gram використовує слово для прогнозування цільового контексту. Як правило, метод skip - gram може мати кращу продуктивність в порівнянні з методом CBOW, оскільки він може захопити два семантичні значення для одного слова.

Окрім word2vec є ще техніка векторних представлень, наприклад, така як GloVe, яка прагне розв'язати проблему захопленням значення одного word embedding із структурою усього осяжного корпусу. Щоб зробити це, модель шукає глобальні збіги числа слів і використовує статистику, мінімізує середньоквадратичне відхилення, видає простір вектору слова з субструктурою. Така схема достатньою мірою дозволяє ототожнювати схожість слова з векторною відстанню.

Перелічені вище алгоритми дозволяють дослідникові провести якісний і що найголовніше автоматизований контент-аналіз текстових даних. Програмні реалізації цих алгоритмів доступні на мові Python в таких пакетах як scikit, R, Gensim і тд.

Запропонований метод контент-аналізу новинних сайтів за кримінальною та географічною ознакою. У рамках нашого дослідження, ми аналізуємо потоки новин, які представлені у вигляді корпусів текстів, що є матеріалами новин з сайтів Казахстану (patrul.kz., inform.kz., azattyq.org), України (tsn.ua., criminal.tv., 24tv.ua), Британії (news.sky.com) і США (foxnews.com, cnn.com), присвячених кримінальній тематиці. Усі перераховані сайти є достовірними джерелами інформації і містять дані про протиправні (кримінальні) дії тієї або іншої країни.

Для автоматичного збору даних було спроектовано додаток на мові Python, яке дозволило рахувати статті з вказаних сайтів новин і зберігати їх у БД [14].

id	head	url	text
5657	В Львове :	https://tsn.ua/ru/ukrayina/v-lvove-zy	Во Львове правоохранители задержа
5658	В Австрал	https://tsn.ua/ru/svit/v-avstralii-polici	Австралийская полиция изъяла круп
5659	Луценко г	https://tsn.ua/ru/politika/lucenko-pri	Генеральный прокурор Украины Юр
5660	Празднов	https://tsn.ua/ru/ukrayina/prazdnovar	Украинские правоохранители зареги
5661	15 тысяч	https://tsn.ua/ru/ukrayina/15-tysyach	За пять лет с 2013 по 2018 год в Украи
5662	lazyload c	https://tsn.ua/ru/ukrayina/iz-ukrainy-	Правоохранители Хмельницкой обла
5663	lazyload c	https://tsn.ua/ru/ukrayina/izbil-i-pyta	В селе Армашовка Ширяевского рай
5664	Минздрав	https://tsn.ua/ru/ukrayina/minzdrav-p	Гослекслужба на просьбу Министер
5665	Количест	https://tsn.ua/ru/ukrayina/kolichestvc	В течение 26 дней предвыборной кам
5666	В Ивано-(	https://tsn.ua/ru/ukrayina/v-ivano-fra	В Ивано-Франковске в воскресенье,
5667	На Днепр	https://tsn.ua/ru/ukrayina/na-dnepro	На в Криничанск

Рисунок 1 – Приклад фрагменту БД, у якій зберігаються зібрані тексти з сайтів новин

На сьогодні об'єм корпусу складає 10 500 текстів, що безперервно поповнюється. Тексти з британських і американських сайтів представлені англійською мовою, а з казахських і українських російською мовою.

Тексти корпусу розподілені по 4 підкорпусам, що відповідають за деяку з країн. Також були розроблені два окремі корпуси (один з усіма російськими текстами, а інший з англійськими) для створення тематичних словників кримінально-забарвленої лексики для двох мов.

В процесі контент-аналізу усі тексти новин були проаналізовані по видах злочинів. Згідно з оцінкою експертів, в нашому корпусі текстів переважають такі види злочинів як:

- вбивство,
- викрадення або пропажа людини,
- ДТП,
- злочини, пов'язані з неповнолітніми,
- хуліганство,
- шахрайство,
- крадіжка (грабіж),
- наркотики,
- викрадення авто.

По перерахованих категоріях був сформований словник кримінально-забарвлених термінів у форматі XML по трьох частинах мови: іменники, дієслова і прикметники. Кожен тег <term> виділяє словникову статтю, <lemma> виділяє слово, що має свою частину мови – <POS>, категорію злочину – <class> і список синонімічних слів – <sin>.

```
<term>
        <lemma>езда</lemma>
        <POS> NOUN </POS>
        <class> ДТП </class>
        <synset>гонка, движение, передвижение </synset>

</term>
<term>
        <lemma>нарушение</lemma>
        <POS> NOUN </POS>
        <class> ДТП </class>
        <synset>несоблюдение, отступление, переступание</synset>
</term>
```

Рисунок 2 – Приклад словарної статті у категорії «ДТП»

Ми виділяємо дві задачі здійсненого контент-аналізу:

- ранжирування документів по мірі тематичної схожості. В даному випадку ми розглядаємо різні види кримінальних злочинів;
- визначення тематичної спрямованості широко-відомих сайтів новин для побудови "кримінальної картини" тієї або іншої країни і розгляду залежності між злочинами, їх географічним розташуванням і часом. Ми плануємо зіставити рівні злочинності чотирьох країн (Україна, Казахстан, Британія і США) і классифікувати злочини за часовими шкалами.

Для вирішення подібних завдань і для роботи з векторними моделями слів (такими як Word2Vec, FastText і т. д.) а також для створення так званих тематичних моделей текстів був вибраний пакет Gensim - бібліотека обробки природної мови, що призначена для "Тематичного моделювання". Це процес побудови тематичної моделі колекції текстових документів, яка визначає, до яких тем відноситься кожен документ колекції. Алгоритм побудови тематичної моделі отримує на вході колекцію текстових документів, а на виході для кожного документа видається числовий вектор, складений з оцінок міри приналежності цього документу кожній з тем. Розмірність цього вектору, рівна числу тем, що може бути задана на вході або визначатися моделлю автоматично.

За допомогою цієї бібліотеки планується проаналізувати тексти корпусів по кримінально-забарвленій лексиці (з використанням ручного словника, створеного на цьому етапі дослідження), визначити тематичну спрямованість тих або інших сайтів новин для подальшого визначення рівня злочинності 4 країн (Британія, США, Україна і Казахстан) в часовій залежності.

Висновки та перспективи подальших досліджень. Важливим і невід'ємним елементом сучасного суспільства в останні роки стали засоби масової інформації. ЗМІ проникають

в усі області людської діяльності і захоплюють увесь інформаційний простір. Сучасне суспільство приймає засоби масової інформації як важливе і значне джерело отримання знань про світ і про те, що у ньому відбувається. Саме це робить їх головним учасником процесу формування громадської думки, культури і світогляду людей.

У епоху розвитку способів комунікації, ЗМІ є "соціальним інструментом, що забезпечує взаємодію в текстовому форматі з метою модифікації картини світу індивіда". Саме тому виникла необхідність у вивченні, глибокому аналізі і осмисленні ролі мови ЗМІ.

Одним з сучасних і ефективних методів для такого вивчення є контент-аналіз або так званий аналіз змісту. Він є процедурою аналізу різних видів текстів (вербальних і візуальних) і дозволяє аналізувати те що лежить між комунікатором і аудиторією. Контент-аналіз результативніший при завданні аналізу засобів масової інформації чим, наприклад, опитний метод оскільки він не втручається в те що вивчає, а досліджує великі об'єми текстової інформації, виділяючи в ній основні аспекти, які не видно неозброєним оком.

Контент-аналіз може бути використаний для абсолютно різних завдань дослідження, головне визначитися з його спрямованістю. Саме тому для нашого дослідження, яке стосується вивчення інформаційного веб-простору 4 держав із заданої тематики, такий вид аналізу призведе до найбільш точного і місткого результату.

З його допомогою проаналізовано корпуси текстів, які містять статті з провідних сайтів новин таких країн як: Великобританія, Україна, США і Казахстан. Метою аналізу видається визначення кримінальної картини країни, за допомогою виділення кримінально-забарвленої лексики сайтів новин, а також подальший аналіз результатів і його класифікація за часовим, географічним і тематичним (рівень різних видів злочинів) критерієм.

Було розроблено словник, криминально-детермінованих термінів, який було оформлено у форматі XML, російською мовою. Словник впорядковано за частинами мови і видами злочинів.

На наступному етапі дослідження передбачається розробка англомовного словника кримінально-забарвлених термінів, за допомогою якого буде проведено контент-аналіз новинного веб-простору таких країн як США та Великобританія, що присвячений класифікації за часовим, тематичним та географічним критерієм. Після чого планується об’єднати результати та провести порівняльний аналіз отриманих даних.

ЛІТЕРАТУРА:

1. Steven E. Stemler (2001), “An Overview of Content Analysis”, Practical Assessment, Research & Evaluation, 7(17)] Available online: http://PAREonline.net/getvn.asp?v=7&n=17.

2. Klaus H Krippendorff (2013), “Content Analysis - 3rd Edition: an Introduction to Its Methodology”, Thousand Oaks: SAGE Publications, Inc, 422 p.

3. Steven E. Stemler (2015), “Emerging Trends in Content Analysis”, Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource, pp. 1-14.

4. Daniel Riffe, Stephen Lacy and Frederick Fico (2014), “Analyzing Media Messages Using Quantitative Content Analysis in Research”, Routledge, 200 p.

5. Walter L. Perry, Brian McInnis, Carter C. Price, Susan Smith, John S. Hollywood (2013), “Predictive Policing Forecasting Crime for Law Enforcement”, RAND Corporation, 4 p.

6. Chelsea H. Smith, Adam Lankford (2016), “The linguistics of terror: a content analysis of suicide notes and martyr manifestos”, the Department of Criminal Justice in the Graduate School of The University of Alabama, 67 p.

7. Hossein Hashemnezhad (2015), “Qualitative Content Analysis Research: A Review Article”, Journal of ELT and Applied Linguistics (JELTAL), pp. 54-62.

8. Berg, B.L. (2001), “Qualitative Research, Message for the Social Sciences”, 4th Edition, Allin and Bacon, pp. 15-35.

9. Mayring P. (2000), “Qualitative Content Analysis”, Forum Qualitative Sozialforschung, 10 p.

10. Hsiu-Fang Hsieh, Sarah E Shannon (2005), “Three Approaches to Qualitative Content Analysis”, Qualitative Health Research, pp. 1277-1288.

11. H. Andrew Schwartz, Lyle H. Ungar (2015), “Data-Driven Content Analysis of Social Media”, The Annals of the American Academy of Political and Social Science, pp. 78-94.

12. Hea-Jin Kim, Yoo Kyung Jeong and Min Song (2015), " Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news", Journal of Information Science, pp. 2-20.

13. Chen, S. (2016). "Language and ecology: A content analysis of ecolinguistics as an emerging research field", Ampersand 3, pp. 108-116.

14. Khairova N., Kolesnyk A., Mamyrbayev O. and Mukhsina K., (2019), "The Aligned Kazakh-Russian Parallel Corpus Focused on the Criminal Theme." COLINS, pp. 116-125.

Колесник А.С., д.т.н., проф. Хайрова Н.Ф.

ИСПОЛЬЗОВАНИЕ МЕТОДОВ КАЧЕСТВЕННОГО КОНТЕНТ-АНАЛИЗА ДЛЯ ИССЛЕДОВАНИЯ ВЕРОЯТНОСТНОГО РАСПРЕДЕЛЕНИЯ УГОЛОВНО ЗНАЧИМОЙ ИНФОРМАЦИИ НА НОВОСТНЫХ САЙТАХ

Сегодня, в век развития информационного общества, СМИ играют роль мощного инструмента формирования общественного мнения и влияния на него. Сейчас всю информацию можно найти в текстовом виде в интернете, в том числе и через ресурсы социальных медиа. Для анализа такого рода данных лучше всего применять актуальную на сегодняшний день информационную технологию контент-анализа, который изучает документы в их социальном контексте и используется при исследовании тематической направленности средств массовой информации. Благодаря развитию методов контент-анализа появилась возможность автоматического исследования содержания текстов, их эффективности и оценки влияния на общество. В данном исследовании приводится анализ существующих подходов, методов и инструментов контент-анализа, а также обосновывается актуальность исследования возможностей использования широкого спектра лингвистических категорий для качественного контент-анализа. Рассматриваются концептуальные возможности использования данного вида анализа в современных лингвистических и социальных исследованиях. В статье показано использование методов качественного контент-анализа, базирующихся на использовании подходов machine learning и разработанного трехъязычного словаря криминально окрашенных терминов, который является одним из основных инструментов исследования распределения криминально значимой информации новостных сайтов веб-медиа по географическим, временным признакам и категориям преступлений. В исследовании также предлагаются базовые основы разработки информационной технологии контент-анализа новостного веб-пространства определенных географических регионов во временной зависимости по заданной тематике, а именно криминальной картине региона. В качестве экспериментального корпуса рассматриваются тексты, собранные автоматически, с помощью разработанного программного продукта, с новостных сайтов Казахстана, Украины, Великобритании и США.

Ключевые слова: качественный контент-анализ; новостные веб-сайты; криминально значимая информация; средства массовой информации; казахский, украинский и английский текстовые корпуса, словарь криминально-окрашенных терминов; методы machine learning

Kolesnyk Anastasiia, doctor of technical sciences, prof. Khairova Nina

THE USE OF QUALITY CONTENT ANALYSIS METHODS TO INVESTIGATE THE PROBABILITY OF CRIMINALLY SIGNIFICANT INFORMATION ON NEW WEB SITES

Today, in the age of the information society, the media play a powerful role in shaping and influencing public opinion. Accordingly, it is a social phenomenon, which affects the point of view of the society. Now all information can be found in text form on the Internet, especially with the help of social media resources. Implementation of such relevant information technology as content analysis is the best way to analyze such kind of data. This method studies documents in their social context and it is used when examining the thematic orientation of the media. At the same time, thanks to the development of methods of content analysis, now it is possible to automatically study the content of different texts, their effectiveness and assess the impact on society. This study analyses existing approaches, methods and tools for content analysis and justifies the relevance of exploring the use of a wide range of linguistic categories for qualitative content analysis. Conceptual possibilities of using this type of analysis in modern linguistic and social research are also considered. The article shows the use of qualitative content analysis methods, based on the use of machine learning approaches and the developed three-language dictionary of criminally colored terms, which is one of the main tools for examining the distribution of criminally significant

information of web media news sites by geographical, time characteristics and categories of crime. In this study, we also offer the bases of the development of content analysis information technology of news web space of certain geographical regions that are analyzed in time dependence on the given topic, namely criminal picture of the region. The texts of news sites of Kazakhstan, Ukraine, Great Britain and the USA were assembled automatically using the developed software product. They are considered as an experimental corpus.

Keywords: qualitative content analysis; quantitative content analysis; News websites; Criminal information; mass media; Kazakh, Ukrainian and English text corpora; dictionary of criminal-colored terms; Machine learning methods