Міністерство освіти і науки України
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»
Кафедра інформатики

Автоматизований аналіз мотиваційних листів на основі сентимент аналізу

Текстова частина до курсової роботи
за спеціальністю «Інженерія програмного забезпечення»

Керівник курсової роботи
доцент, к.н. Ковалюк Т. В.
_____
(підпис)
“____” _____ 2020 р.
Виконав студент Рожко Р. О.
“____” _____ 2020 р.

Київ 2020

Зміст

Анотація

Рожко Р. О. Автоматизований аналіз мотиваційних листів на основі сентимент-аналізу.

У курсовій роботі була описана проблема обробки текстів природньої мови, описані способи такої обробки, особливо сконцентровано увагу на сентимент-аналізі; проблеми та способи його застосування; описані алгоритм роботи класифікаторів текстових даних; процес розробки та сама система аналізу.

Робота складається з трьох розділів.

Перший розділ присвячений аналізу сучасного стану вирішення проблеми, обґрунтуванню необхідності виконання роботи, призначення курсової роботи, сучасний стан проблем сентимент-аналізу та спосіб його застосування.

Другий розділ розповідає про процес обробки текстових даних для їх подальшого аналізу, принципам роботи класифікатора тексту на основі машинного навчання.

Третій розділ присвячений огляду системи, її основних параметрів та подальших кроків для вдосконалення її роботи.

Вступ

Темою цієї робити є «Автоматизований аналіз мотиваційних листів для вступу у ВНЗ на основі сентимент аналізу». Сьогодні мотиваційний лист – це не просто формальність чи ввічливість. Добре написаний мотиваційний лист – це можливість показати читачеві свої: особистісні якості, ваш вклад у розвиток університету та прагнення здобувати освіту у вибраному вищому навчальному закладі. Особливо актуальною ця тема є сьогодні в Україні. У січні 2020 року в силу вступили зміни до закону про вищу освіту, який передбачає, що заклади вищої освіти прийматимуть мотиваційні листи, за які абітурієнт зможе отримати додаткові бали під час вступу.

Одним з завдань програмування є автоматизація процесів. Отже метою цієї курсової роботи є розробка програмного забезпечення яке б дозволило за допомогою програмних засобів оцінити та проаналізувати мотиваційні листи для вступу у вищий навчальний заклад. У цій роботі я розгляну сучасні методи аналізу текстових документів, проблеми комп’ютерного аналізу та синтезу природної мови, принципи роботи штучного інтелекту та особливо сконцентруюся на аналізі тональності тексту – клас методів контент-аналізу в комп’ютерній лінгвістиці, призначений для автоматизованого виявлення в текстах емоційно забарвленої лексики і емоційної оцінки авторів (думок) по відношенню до об'єктів, мова про які йде в тексті. У ході цієї роботи будуть розглянуті способи оцінки мотивації абітурієнтів, розроблені точні критерії аналізу, такі які вклад потенційного студента в навчальний процес вищого навчального закладу, бажання навчатися за обраною спеціальністю та інші. Завершальною частиною цієї курсової роботи буде огляд сучасних інструментів для роботи з сентимент

аналізом та процесу розробки системи аналізу мотиваційних листів та її роботоздатність.

Результати цієї курсової роботи можуть принести практичне значення нашому університету. Дослідження та система розроблена у ході цієї роботи можуть бути початковим етапом у створенні програмного продукту, який би спростив відбір мотиваційних листів вступників та їх подальшою обробкою у межах університету.

1. Проблеми аналізу тональності тексту

1.1. Визначення сентимент-аналізу

Сентимент-аналіз або аналіз тональності тексту (англ. Sentiment analysis, англ. Opinion mining) відноситься до застосування обробки природньої мови, комп'ютерної лінгвістики та аналітики тексту для виявлення та класифікації суб'єктивних думок у вхідних джерелах. Взагалі аналіз тональності має на меті визначати ставлення автора до якоїсь теми або загальної контекстуальної полярності тексту. Ставлення може бути його судженням або оцінкою, афективним станом (тобто емоційним станом автора під час написання) або наміченим емоційним спілкуванням (тобто емоційним впливом, який автор хоче мати на читача).

Аналіз тональності тексту включає багато завдань, таких як визначення тональності, класифікація настроїв в тексті на такі категорії як «позитивні» чи «негативні» або «нейтральні», класифікація суб'єктивного ставлення автора, узагальнення думок. Він спрямований на аналіз настроїв, ставлення, емоційних думок до таких речей, як товари, особистості, теми, організації чи послуги.

Як правило, сентимент-аналіз класифікує текстові вирази у вихідних джерелах на два типи: факти – об'єктивні висловлювання про сутності, події та їхні атрибути, наприклад, «я вчора купив футболку»; та думки – суб'єктивні висловлювання настроїв, емоцій, оцінки або почуття до сутності, наприклад, «мені дуже подобається її колір». Слід зазначити, що не всі суб'єктивні речення містять думки і не всі об'єктивні речення не містять думок. Тому для аналізу тональності важливо виявити та витягнути факти та думки з вихідних матеріалів. Однак, на жаль, цього важко досягти.

Термін сентимент-аналіз імовірно вперше з’явився в 2003 році[2]. Однак дослідження тональності та думок з’явилося ще раніше.

Хоча лінгвістика та обробка природних мов мають довгу історію, проте мало досліджень було зроблено щодо думок та тональності людей до 2000 року. З тих пір ця сфера почалася дуже активно досліджуватися. Для цього є кілька причин. По-перше, вона має широке застосування майже у кожній доменній області. Аналіз тональності став активно застосовуватися у комерційних застосунках, що дало сильний поштовх досліджень. По-друге, це відкрило велику кількість дослідницьких викликів, які раніше не вивчалися.

1.2. Застосування сентимент-аналізу

Думки є центральними у майже всій діяльності людини, оскільки вони є ключовими чинниками нашої поведінки. Щоразу, коли нам потрібно прийняти рішення, ми хочемо знати думки інших. У реальному світі підприємства та організації завжди хочуть дізнатися думки споживачів чи громадськості щодо своїх товарів та послуг. Споживачі також хочуть дізнатися думку існуючих користувачів товару перед його придбанням, а також думки інших щодо політичних кандидатів, перш ніж приймати рішення про голосування на політичних виборах. Раніше, коли комусь потрібні були відгуки людей, він запитував друзів та родину. Коли організації чи бізнесу потрібні відгуки та настрої громадськості чи споживачів, вони проводять опитування, та створюють фокус-групи. Продаж відгуків громадськості та споживачів вже давно є величезним бізнесом для маркетингу та політичних кампаній.

Із стрімким зростанням соціальних медіа (наприклад, огляди, дискусії на форумах, блоги, мікро-блоги, коментарі та публікації на сайтах

соціальних мереж) у мережі інтернет люди та організації все частіше використовують вміст у цих ЗМІ для прийняття рішень. . Сьогодні, якщо ви хочете придбати споживчий продукт, більше не обмежуйтеся запитанням у своїх друзів та родини, оскільки в Інтернеті існує багато оглядів та дискусій на громадських форумах. Для організації, можливо, більше не потрібно проводити опитування та фокус-груп для збору громадської думки, оскільки існує маса такої інформації, яка є загальнодоступною. Однак пошук та моніторинг сайтів в Інтернеті та обробка інформації, що міститься в них, залишається важким завданням. Кожен сайт зазвичай містить величезний обсяг відгуків користувачів, які не завжди легко розшифрувати у довгих блогах та публікаціях на форумах. Пересічній людині буде складно визначити відповідні сайти, витягнути та узагальнити інформацію на них. Таким чином, необхідні автоматизовані системи аналізу тональності.

Останніми роками ми були свідками того, що суб'єктивні публікації в соціальних медіа допомагали переробити бізнес, а також змінити громадські настрої та емоції, які сильно вплинули на нашу суспільну та політичну систему. Таким чином, це стало необхідністю збирати та вивчати думки в Інтернеті. Зрозуміло, Інтернет не єдине джерело таких даних, багато організацій також мають свої внутрішні дані, наприклад, відгуки клієнтів, зібрані з електронних листів та кол-центрів, або результати опитувань, проведених компаніями.

Завдяки цим додаткам, промислова діяльність процвітала в останні роки. Програми аналізу тональності поширилися майже на всі можливі сфери - від споживчих товарів, послуг, охорони здоров'я та фінансових послуг до соціальних заходів та політичних виборів. Багато великих корпорацій також створили власні внутрішні системи, наприклад, Microsoft, Google,

HewlettPackard, SAP та SAS. Ці практичні додатки та промислові інтереси дали сильну мотивації для дослідження аналізу настроїв.

1.3 Рівні сентимент-аналізу

Зараз я коротко ознайомлююся з основними проблемами дослідження на основі рівня деталізації існуючих досліджень. В цілому аналіз настроїв досліджувався в основному на трьох рівнях.

1.3.1 Рівень документа

Завдання на цьому рівні - класифікувати, чи виражає цілий документ позитивні чи негативні настрої. Наприклад, даючи відгук про товар, система визначає чи є відгук в цілому позитивний чи негативний щодо товару. Це завдання загальновідоме як класифікація настроїв на рівні документа. Цей рівень аналізу передбачає, що кожен документ виражає думки щодо однієї сутності, наприклад, одного товару чи послуги. Таким чином, це не застосовується до документів, які оцінюють або порівнюють кілька об'єктів.

1.3.2 Рівень речення

Завдання на цьому рівні переходить до речень і визначає, чи висловлює кожне речення позитивну, негативну чи нейтральну думку. Нейтральний зазвичай означає відсутність оцінки. Цей рівень аналізу тісно пов'язаний із класифікацією суб'єктивності, яка відрізняє речення (називаються об’єктивними реченнями), які виражають фактичну інформацію від речень (називаються суб'єктивними реченнями), які виражають суб'єктивні погляди та думки. Однак слід зазначити, що суб'єктивність не є рівнозначною

настрою, оскільки багато об'єктивних пропозицій можуть наводити на певні погляди.

1.3.3 Рівень сутності та аспекту

Як аналіз рівня документа, так і аналіз рівня речень не виявляють, що саме людям подобалось і що не подобалось. Аспектний рівень виконує тонкозернистий аналіз. Замість того, щоб дивитись на мовні конструкції (документи, абзаци, речення чи фрази), рівень аспекту безпосередньо дивиться на саму думку. Він заснований на ідеї, що погляд людини на ту чи іншу сутність складається з настроїв (позитивних чи негативних) та цільових думок. Думка, не визначена її мета, є обмеженою. Усвідомлення важливості цілей думок також допомагає нам краще зрозуміти проблему аналізу тональності. Наприклад, хоча речення "хоча сервіс не такий хороший, я все одно люблю цей ресторан" явно має позитивний тон, але не можна сказати, що це речення є повністю позитивним. Насправді речення є позитивним щодо ресторану, але негативно щодо його обслуговування. У багатьох додатках цілі опитування описуються суб'єктами та / або їх різними аспектами. Таким чином, метою цього рівня аналізу є виявлення настроїв щодо сутності та / або їх аспектів. Наприклад, речення "Якість дзвінків iPhone хороша, але термін її акумулятора короткий" оцінює два аспекти, як якість дзвінка, так і термін служби акумулятора, iPhone (сутність). Настрої на якість дзвінків iPhone позитивні, але настрої на час автономної роботи негативні. Якість виклику та час автономної роботи iPhone - цільові думки. На основі такого рівня аналізу може бути створений структурований підсумок думок про сутності та їх аспекти, який перетворює неструктурований текст на структуровані дані і може бути використаний для всіх видів якісного та кількісного аналізу. І класифікація

рівня документа, і рівня речень вже є дуже складним завданням. Аспект-рівень ще складніше. Він складається з декількох підпроблем, про які ми поговоримо у наступних розділах.

1.4 Сентимент-лексикон

Найважливішими показниками тональності є сентимент-слова (англ. sentiment words). Це слова, які зазвичай використовуються для вираження позитивних чи негативних настроїв. Наприклад, хороший, чудовий та дивовижний - це позитивні сентиментальні слова, а поганий та жахливий - слова негативної тональності. Крім окремих слів, існують також фрази та ідіоми. Сентимент-слова та фрази сприяють аналізу настроїв із зрозумілих причин. Перелік таких слів і фраз називається сентимент-лексиконом. Протягом багатьох років дослідники розробили численні алгоритми для складання таких лексиконів. Хоча слова та фрази сентименту важливі для аналізу настроїв, лише їх використання далеко не достатньо. Проблема набагато складніша. Іншими словами, можна сказати, що сентимент-лексикон необхідний, але недостатній для аналізу настроїв. Нижче наведено кілька проблем пов’язаних з цим:

1. Слова позитивної чи негативної тональності можуть мати протилежні значенні в залежності від області їхнього застосування.
2. Речення, що містять сентимент-слова, можуть не виражати жодних поглядів. Це явище найчастіше трапляється в питальних реченнях та підрядних реченнях умови, наприклад, «Якщо я знайду хороший пилосос у магазині, то я його куплю» та «Чи можете мені підказати чи цей телефон хороший».

Обидва речення містять сентимент-слово «хороший», але жодне не виражає позитивної чи негативної думки до будь-якого конкретного товару. Однак не всі питальні речення чи підрядні речення умови не висловлюють думок.

3. Саркастичні речення зі сентимент-словами або без них важко розібрати. Сарказм не так часто зустрічається в оглядах споживачів щодо товару або послуги, але часто є дуже поширеним у політичних дискусіях або мотиваційних листах, через що такі матеріли важко аналізувати.
4. Багато речень можуть не містити в собі сентимент-слів, але вони можуть висловлювати якісь думки. Багато з цих речень – це фактично об’єктивні речення, які використовуються для вираження деякої фактичної інформації. Наприклад, «Ця пральна машина використовує багато води» несе негативний характер, оскільки пральна машина використовує занадто багато ресурсів.

1.5 Визначення думки

Думка складається з двох компонентів: суб’єкта g, та тональності (настрою) s до цього суб’єкту,

(g, s),

де g може бути будь-яким суб'єктом або аспектом суб'єкта, щодо якого висловлена думка, і s - позитивний, негативний або нейтральний настрій, або числовий бал, що виражає силу / інтенсивність настроїв (наприклад, від 1 до 5). Позитивні, негативні та нейтральні називаються сентимент орієнтацією (або полярністю).

Також думку можна висловити як функцію з чотирьох аргументів:

(g, s, h, t),

де g – суб’єктом настроїв, s - настрої щодо суб’єкта, h - власник оцінки і t - час, коли думка була висловлена.

Це визначення може бути непросто використовувати на практиці, особливо в галузі оглядів товарів, послуг та брендів в Інтернеті, оскільки повний опис суб’єкта може бути складним і навіть може міститися не в одному реченні. На практиці суб’єкт часто може бути розкладений та описаний структуровано з різними рівнями, що значно полегшує як видобуток думок, так і пізніше використання результатів видобутої думки. Наприклад, "якість зображення Canon G12" може бути розкладено на суб’єкт і атрибут суб’єкта і представити у вигляді пари,

(Canon-G12, якість зображення),

Визначимо термін сутність для позначення цільового суб’єкта, який був оцінений. Сутність можна визначити наступним чином. Суб'єкт e - це продукт, послуга, тема, питання, особа, організація чи подія. Він описується парою, e: (T, W), де T - ієрархія частин, підрозділів тощо, а W - сукупність атрибутів e. Кожна частина або підрозділ також має свій набір атрибутів. Це визначення по суті описує ієрархічну декомпозицію сутності на основі часткового відношення. Кореневий вузол - це назва сутності. Усі інші вузли - це частини та підрозділи тощо. Думка може бути висловлена на будь-якому вузлі та будь-якому атрибуті вузла.

Ця сутність як ієрархія будь-якої кількості рівнів потребує вкладеного відношення для її представлення, що часто є занадто складним для застосунків. Основна причина полягає в тому, що оскільки обробка природньої мови є дуже складним завданням, розпізнавати частини та атрибути суб'єкта господарювання на різних рівнях деталей є надзвичайно

важко. Більшість застосунків також не потребують такого складного аналізу. Таким чином, ми спрощуємо ієрархію до двох рівнів і використовуємо термінові аспекти для позначення обох частин і атрибутів. У спрощеному дереві кореневий вузол все ще є самою сутністю, але вузли другого рівня - це різні аспекти сутності. Цей спрощений фреймворк - це те, що зазвичай використовується в практичних системах аналізу настроїв.

Зауважимо, що в дослідницькій літературі сутності також називають об'єктами, а аспекти також називають ознаками. Однак тут функції можна сплутати з функціями, які використовуються в машинному навчанні, де функція означає атрибут даних. Щоб уникнути плутанини, останніми роками частіше використовують термін аспект. Зауважте, що деякі дослідники також використовують терміни фасети, атрибути та теми, а в конкретних програмах сутності та аспекти також можуть називатися іншими іменами, заснованими на домовленостях домену додатків. Розклавши суб’єкт думки, ми можемо переосмислити думку.

Думка як функція з п’яти аргументів:

$(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$,

де e_i - назва сутності, a_{ij} - це аспект e_i, s_{ijkl} - це настрої щодо аспекту a_{ij} сутності e_i, h_k - власник оцінки, а t_l - час, коли думка була висловлена h_k. Настрій s_{ijkl} є позитивним, негативним чи нейтральним або вираженим з різними рівнями сили / інтенсивності, наприклад, від 1 до 5 зірок, як використовується більшість сайтів відгуків в інтернеті. Коли думка стосується самої сутності в цілому, для її позначення використовується спеціальний аспект GENERAL. Тут e_i та a_{ij} разом представляють суб’єкт.

Деякі важливі зауваження щодо цього визначення:

1. У визначенні думка s_{ijkl} обов’язково повинна бути надана власником думки h_k про аспект a_{ij} сутності e_i в момент t_l. Будь-яка невідповідність - це помилка.
2. Усі п'ять компонентів є важливими. Відсутність будь-якого з них загалом проблематична. Наприклад, якщо у нас немає компонента часу, ми не зможемо проаналізувати думки щодо суб'єкта господарювання відповідно до часу, що часто дуже важливо на практиці, оскільки думка два роки тому та думка вчора не є однаковою. Без власника думки також проблематично. Наприклад, у реченні "міського голови люблять люди міста, але його критикує державна влада", чітко важливими є два носії думки, "люди у місті" та "державна влада".
3. Визначення охоплює більшість, але не всі можливі грані смислового значення думки, яке може бути доволі складним. Він також не охоплює контекст думки, наприклад, "Цей автомобіль занадто малий для високої людини", що не говорить про те, що машина занадто мала для всіх. "Високий чоловік" - це контекст. Зауважимо також, що в початковому визначенні сутності це ієрархія частин, підрозділів тощо. Кожна частина може мати свій набір атрибутів. Завдяки спрощенню, представлення може призвести до втрати інформації. Наприклад, "чорнило" - це частина / компонент принтера. В огляді принтера написано: «Чорнило цього принтера дороге». Це не говорить про те, що принтер дорогий (що вказує на аспектну ціну). Якщо жоден атрибут чорнила не важливий, це речення просто дає негативну думку до чорнила, що є аспектом

сутності принтера. Однак, якщо також хочеться вивчити думки щодо різних аспектів чорнила, наприклад, ціни та якості, до чорнила потрібно ставитися як до окремої сутності. Звичайно, концептуально ми також можемо розширити представлення суб'єктів думок, використовуючи вкладене відношення. Незважаючи на обмеження, визначення охоплює істотну інформацію думки, достатню для більшості застосунків. Як ми згадували вище, занадто складне визначення може зробити проблему надзвичайно важкою.

4. Це визначення забезпечує основу для перетворення неструктурованого тексту в структуровані дані. Функція вище - це в основному схема бази даних, на основі якої витягнуті думки можна помістити в таблицю бази даних. Тоді багатий набір якісного, кількісного та тенденційного аналізу думок можна виконати за допомогою всього набору систем управління базами даних (СУБД) та інструментів OLAP.

1.6 Завдання сентимент-аналізу

За допомогою визначення, ми можемо представити об'єктивні та ключові завдання аналізу настроїв.

Мета аналізу настроїв: маючи документ d, визначити усі думки (e_i, a_{ij}, s_{ijkl}, h_k, t_l) у d.

Основні завдання отримані з 5 компонентів функції. Перший компонент - це сутність. Тобто нам потрібно витягти сутності. Завдання аналогічна розпізнаванню названої сутності (NER) при вилученні інформації. Таким чином, саме видобуток є проблемою. Після видобутку нам також

потрібно класифікувати видобуті сутності. У тексті природньою мовою люди часто пишуть одне й те саме по-різному. Наприклад, Motorola може писатися як Mot, Moto та Motorola. Нам потрібно визнати, що всі вони відносяться до однієї сутності.

Категорія сутності являє собою унікальну сутність, тоді як вираз сутності - це фактичне слово чи фраза, що з’являється в тексті, що вказує на категорію сутності. Кожна категорія сутності (або просто сутність) повинна мати унікальну назву. Процес групування виразів сутності в категорії сутностей називається категоризацією сутності.

Зараз ми розглянемо аспекти сутностей. Проблема в основному така ж, як і для сутностей. Наприклад, зображення, зображення та фото - це той самий аспект для камер. Таким чином, нам потрібно витягнути вирази аспектів і категоризувати їх.

Категорія аспекту сутності являє собою унікальний аспект сутності, тоді як вираз аспекту - це фактичне слово або фраза, яка з’являється в тексті, що вказує на категорію аспекту. Кожна категорія аспектів (або просто аспект) також повинна мати унікальну назву. Процес групування виразів аспектів у категорії аспектів (аспектів) називається категоризацією аспектів.

Виразними аспекту зазвичай є іменники та іменникові фрази, але можуть бути також дієслова, дієприслівники, прикметники та прислівники. Вирази аспекту, які є іменниками та іменниковими словосполученнями, називаються явними аспектними виразами. Вирази аспекту, які не є іменниками чи іменниковими словосполученнями, називаються неявними аспектними виразами.

Наприклад, "дорога" - це неявний вираз аспекту в "Ця камера дорога". Це передбачає аспектну ціну. Багато неявних аспектних виразів - це прикметники та прислівники, які використовуються для опису або

класифікації деяких конкретних аспектів, наприклад, дорогий (ціна) та надійний (надійність). Вони також можуть бути дієсловами, наприклад, "Я можу легко встановити програмне забезпечення". "Встановити" вказує на аспект установки. Неявні аспектні вирази - це не просто прикметники, прислівники, дієслова; вони також можуть бути дуже складними, наприклад, "Цю камеру не легко помістити в кишені пальто." Тут "помістити в кишені пальто" вказує розмір аспекту (та / або форму).

Третім компонентом у визначенні думки є настрої. Це завдання класифікує, чи позитивні, негативні чи нейтральні настрої щодо аспекту. Четвертий та п'ятий компоненти - це власник думки та час відповідно. Вони також повинні бути вилучені та класифіковані як для сутностей та аспектів. Зауважте, що власником думки (також його називають джерелом думки) може бути людина чи організація, яка висловила думку. Для оглядів товарів і блогів, власники думок, як правило, є авторами публікацій. Власники думок важливіші для новин, оскільки вони часто прямо заявляють про особу чи організацію, яка дотримується думки. Однак у деяких випадках виявлення власників думок також може бути важливим у соціальних медіа, наприклад, визначення думок рекламодавців або людей, які цитують рекламні оголошення компаній.

На основі вищесказаного ми можемо визначити модель сутності та модель документа думки.

Модель сутності: Суб'єкт e_i представлений самим собою в цілому і кінцевим набором аспектів $A_i = \{a_{i1}, a_{i2},\ldots, a_{in}\}$. e_i може бути виражений будь-яким з кінцевих наборів його виразів сутності $\{e_{ei1}, e_{ei2},\ldots, e_{eis}\}$. Кожен аспект $aij \in A_i$ сутності e_i може бути виражений будь-яким його кінцевим набором аспектних виразів $\{a_{eij1}, a_{eij2},\ldots, a_{eijm}\}$.

Модель документа думки: Документ d містить думки про сукупність сутностей $\{e_1, e_2,\ldots, e_r\}$ та підмножину їх аспектів із набору власників думок $\{h_1, h_2,\ldots, h_p\}$ у певний конкретний час.

Підсумовуючи, аналіз тональності складається з наступних 6 основних завдань.

Завдання 1 (вилучення та категоризація сутності): Витягнути всі вирази сутності в D, а також категоризувати або згрупувати синонімічні вирази сутності в кластери сутностей (або категорії). Кожен кластер виразів сутності вказує на унікальну сутність e_i.

Завдання 2 (вилучення та категоризація аспектів): Витягнути всі вирази аспектів сутностей та категоризувати ці вирази аспектів на кластери. Кожен кластер виразів аспекту сутності e_i являє собою унікальний аспект a_{ij}.

Завдання 3 (вилучення та категоризація власників думок): Витягнути власників думок з тексту чи структурованих даних та категоризувати їх. Завдання є аналогом вищезазначених двох завдань.

Завдання 4 (вилучення та стандартизація часу): Витягнути час, коли думка була виражена, і стандартизувати різні формати часу. Завдання також є аналогом вищезазначених завдань.

Завдання 5 (категоризація настрою на аспекти): Визначити, чи є думка про a_{ij} аспекту позитивною, негативною чи нейтральною, або призначити числовому рейтингу настроїв аспект.

Завдання 6: Скласти усі думки (e_i, a_{ij}, s_{ijkl}, h_k, t_l), висловлені в документі d на основі результатів вищезазначених завдань. Це завдання, здається, дуже просте, але насправді у багатьох випадках виявляється досить складним.

2. Аналіз мотиваційних листів

Мотиваційний лист - це спосіб пояснити, чому абітурієнт ідеально підходить для навчання у вибраному університету. Це можливість особисто описати мотивацію та досвід, який він отримав, що призвело до прийняття цього рішення. Те, як він напишете мотиваційний лист, може вирішити зарахують чи не зарахують, особливо це стосується університетів з високими стандартами навчання.

Написання мотиваційного листа - це не просто формальність. Приймальні комісії сприймають супровідний лист дуже серйозно, оскільки він відображає прихильність та наміри заявника. Якість листа свідчить про характер, цілі та амбіції. Вимагаючи мотиваційного листа, приймальна комісія надає можливість описати себе коротким документом, в якому потрібно дати деякі важливі та цікаві відомості про абітурієнта.

Написання мотиваційного листа для вступу до університету може бути часозатратним і складним для деяких абітурієнтів, які часто ставлять собі питання про те, як повинен виглядати лист, що він повинен містити та як переконати приймальну комісію, що вони є правильними кандидатами на вступ до університету.

Дані я наведу приблизну структуру мотиваційного листа для вступу у вищий навчальний заклад.

Вступ. У вступному реченні мотиваційного листа завжди має бути зазначено ваша мета, з якою він був написаний; із наступним реченням, яке представляє абітурієнта. Останнє речення може бути в різних формах, наприклад короткий підсумок вашого поточного стану, здобутих навичок, що стосується вибраного напрямку навчання.

Основна частина. У другому абзаці листа слід зосередитись на тому, щоб "продати себе". Описати всі минулі досягнення, академічні виступи та скласти позитивне враження. Якщо обмежений академічний / професійний досвід, потрібно написати свої нематеріальні достоїнства, такі як ставлення до справ, хороші мовні навички та вміння працювати з людьми. Далі треба дати зрозуміти що абітурієнт дослідив навчальну програму, в яку він подається. Описати можливий вклад потенційного студента в навчальний процес вищого навчального закладу, бажання навчатися за обраною спеціальністю.

Висновок. Потрібно узагальнити основні моменти, які були написані раніше. Зробити висновок, відзначивши свій інтерес у здобутті освіти саме в цьому конкретному навчальному закладі.

2.1. Аналіз за допомогою методів обробки природньої мови.

Далі я розгляну кроки, які треба зробити перед застосуванням методів сентимент-аналізу з метою покращення текстових даних для легшої обробки та вилучення сутностей та їх аспектів.

2.1 Препроцесинг

Аналіз тональності тексту є частиною проблеми класифікації тексту, оскільки у нас є текст як вхідний і настрій як вихідний.

Проблема тут полягає в тому, що для людини розуміння тексту та таких понять, як іронія, не є складним. Але для машини це може бути складним завданням. Це відбувається тому, що, хоча поняття тексту для нас дуже зрозуміле, ми не можемо сказати те саме, коли говоримо про комп’ютери.

З цієї причини перед тим, як подавати текст на аналіз, ми повинні попередньо його обробити.

Прибрати зайві розділові знаки та подвійні пропуски. Такі артефакти та помилку можуть завадити розбити документ на лексеми, такі як речення та слова, що є основним матеріалом для аналізу (Рисунок 2.1). Далі будуть приклади коду на мові програмуванні Python, що ілюструють реалізацію кожного з кроків.

```
def remove_space(text)
    text = text.strip()
    return "".join(text)
```

Рисунок 2.1 – Видалення зайвих пропусків у тексті

Перетворення слів у символи нижнього регістра. Такі слова, як "Книга" та "книга" означають те саме, але якщо їх не перетворити на нижній регістр, вони будуть представлені у векторній просторовій моделі двома різними словами (в результаті чого з'являється більше вимірів).

Токенізація. Це процес розщеплення даного тексту на менші частини, які називаються лексемами. Слова, цифри, розділові знаки та інші можна вважати лексемами (Рисунок 2.2). Для цього і наступних прикладів я буду використовувати бібліотеку з відкритим кодом nltk (Natural Language Toolkit).

```
import ntlk
ntlk.download("punkt")

from ntlk.tokenize import word_tokenize

def tokens(text)
    return word_tokenize(text)
```

Рисунок 2.2 – Розбиття тексту на лексеми

Видалення стоп-слів. Стоп-слова - це найпоширеніші слова в мові, як наприклад для англійської: «the», «a», «on», «is», «all». Ці слова не мають важливого значення і зазвичай вилучаються з текстів. Видалити стоп-слова можна за допомогою бібліотеки nltk (Рисунок 3.3).

```
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

stop_words = set(stopwords.words('english'))

def remove_stop_words(text)
    word_tokens = word_tokenize(text)
    filtered_text = [w for w in word_tokens if not w in stop_words]
    return filtered_text
```

Рисунок 2.3 – Видалення стоп-слів

Стеммінг. Це процес скорочення слова до основи шляхом відкидання допоміжних частин, таких як закінчення чи суфікс. Результати стемінгу іноді дуже схожі на визначення кореня слова, але його алгоритми базуються на інших принципах (Рисунок 2.4).

```
from nltk.stem import PorterStemmer
from nltk.tokenize import word_tokenize
stemmer= PorterStemmer()

def stem(text)
    result = []
    word_tokens = word_tokenize(text)
    for word in word_tokens:
        result.append(stemmer.stem(word))
```

Рисунок 2.4 - Стеммінг

Лематизація. Метою лематизації, як і стеммінгу, є зведення флективних форм до загальної основної форми (Рисунок 2.5). На відміну від

стеммінгу, лематизація не просто прибирає зайві символи у слові. Натомість він використовує лексичні бази даних для отримання правильних базових форм слів. Лематизація є кращою перед стеммінг, оскільки лематизація робить морфологічний аналіз слів.

```
import nltk
from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()

def lemmatize(word)
    return lemmatizer.lemmatize(word, pos='n')
```

Рисунок 2.4 – Лематизація

Після попередньої обробки тексту (препроцесинг) результат може бути використаний для більш складних завдань NLP, наприклад, сентимент-аналіз.

2.2 Класифікація

За даними IBM, близько 80% усієї інформації є неструктурованою, при цьому текст є одним із найпоширеніших типів неструктурованих даних. Через безладний характер тексту, аналіз, розуміння, упорядкування та сортування текстових даних є складним та трудомістким.

Класифікацію (або категоризація) тексту можна здійснити двома різними способами: ручною та автоматизованою класифікацією. У першому людський анотатор інтерпретує зміст тексту та відповідно його класифікує. Цей метод зазвичай може забезпечити якісні результати, але це трудомісткий і дорогий. Останній застосовує машинне навчання, природну обробку мови та інші методи автоматичної класифікації тексту швидшим та економічно вигіднішим способом.

В системі автоматизованого аналізу мотиваційних листів, яка є предметом цієї курсової роботи, буду використовувати гібрідний метод класифікації. Гібридна класифікація поєднує базовий класифікатор, який навчається за допомогою машинного навчання, та систему на основі правил, яка використовується для подальшого покращення результатів. Ці гібридні системи можна легко налаштувати, додавши конкретні правила для тих конфліктуючих тегів, які не були правильно змодельовані базовим класифікатором.

Першим кроком для створення класифікатора за допомогою машинного навчання є вилучення сутностей: використовується метод перетворення кожного тексту у числове подання у вигляді вектора. Один з найбільш часто використовуваних підходів - мішок слів, де вектор представляє частоту слова у заздалегідь визначеному словнику слів.

Наприклад, якщо ми визначили наш словник, щоб він містив такі слова {Це, хороший, не, університет, погано, басейн}, і ми хотіли векторизувати текст "Це хороший університет", у нас було б таке векторне подання цього тексту: (1, 1, 0, 1, 0, 0). Я якості такого словника я використовую інструмент з відкритим вихідним кодом, написаний компанією Google, word2vec (Рисунок 2.5).

```
from nltk.tokenize import sent_tokenize, word_tokenize
import warnings

warnings.filterwarnings(action = 'ignore')

import gensim
from gensim.models import Word2Vec

data = []

for i in sent_tokenize(text):
    temp = []

    for j in word_tokenize(i):
        temp.append(j.lower())

    data.append(temp)

model = gensim.models.Word2Vec(data, min_count = 1,
                               size = 100, window = 5)
```

Рисунок 2.5 – Використання бібліотеки word2vec

Потім алгоритм машинного навчання подається з навчальними даними, які складаються з пар наборів функцій (векторів для кожного прикладу тексту) та тегів (наприклад, спорт, політика) для створення моделі класифікації.

Після того, як класифікатор натренований з достатньою кількістю зразків для навчання, модель машинного навчання може почати робити точні прогнози. Цей же екстрактор функцій використовується для перетворення небаченого тексту в набори функцій, які можна подати в класифікаційну модель, щоб отримати прогнозовані теги.

2.2.1 Алгоритм класифікації

Глибоке навчання - це набір алгоритмів і прийомів, натхненних роботою мозку людини. Класифікація тексту отримала користь від недавнього відродження архітектур глибокого навчання завдяки їхньому потенціалу досягти високої точності з меншою потребою в інженерних функціях. Дві

основні архітектури глибокого навчання, що використовуються в класифікації тексту, - це згорткові нейронні мережі (CNN) та періодичні нейронні мережі (RNN).

З одного боку, алгоритми глибокого навчання вимагають значно більшої кількості навчальних даних, ніж традиційні алгоритми машинного навчання, тобто, принаймні, мільйони тегів, що позначаються. З іншого боку, традиційні алгоритми машинного навчання, такі як SVM та NB, досягають певної межі, коли додавання більшої кількості навчальних даних не покращує їх точність. На відміну від цього, класифікатори глибокого навчання продовжують удосконалюватись від більшої кількості тестових даних.

Я використав модель згорткової нейронної мережі для фіксації семантики тексту. Вхід мережі - це документ D, який є послідовністю слів w1, w2. . . wn. Вихід з мережі містить елементи класу. Ми використовуємо p (k | D, θ) для позначення ймовірності того, що документ класу k, де θ - параметри в мережі.

Поєднуємо слово та його контекст, щоб подати слово. Контексти допомагають нам отримати більш точне значення слова. У нашій моделі використовуємо рекуррентну структуру, яка є двонаправленою рекуррентною нейронною мережею для захоплення контекстів. Ми визначаємо $c_l(w_i)$ як лівий контекст слова w_i та $c_r(w_i)$ як правий контекст слова w_i. І $c_l(w_i)$, і $c_r(w_i)$ - щільні вектори з |c| елементи реальної цінності. Лівий контекст $c_l(w_i)$ слова w_i обчислюється за допомогою рівняння (1), де $e(w_{i-1})$ - слово, вбудоване слово w_{i-1}, яке є щільним вектором з |e| елементи реальної цінності. $c_l(w_{i-1})$ - лівий контекст попереднього слова w_{i-1}. Лівий контекст першого слова в будь-якому документі використовує ті ж параметри, що поділяються $c_l(w_1)$. W_l - матриця, яка перетворює прихований шар (контекст) у наступний прихований шар. W_{sl} - це матриця, яка використовується для поєднання семантики поточного слова

з лівим контекстом наступного слова. f - нелінійна функція активації. Правий контекст $c_r(w_i)$ обчислюється аналогічно, як показано в рівнянні (2). Правий контекст останнього слова в документі розділяє параметри $c_r(w_n)$.

$$c_l(w_i) = f(W^{(l)}c_l(w_{i-1}) + W^{(sl)}e(w_{i-1})) \quad (1)$$

$$c_r(w_i) = f\left(W^{(r)}c_r(w_{i+1}) + W^{(sr)}e(w_{i-1})\right) \quad (2)$$

Як показано в рівняннях (1) та (2), контекстний вектор фіксує семантику всіх лівих та правобічних контекстів. Потім ми визначаємо подання слова w_i у рівнянні (3), яке є конкатенацією лівого контекстного вектора $c_l(w_i)$, словом, що вбудовує $e(w_i)$, і правою частиною контексту вектор $c_r(w_i)$. Таким чином, використовуючи цю контекстну інформацію, наша модель, можливо, зможе краще розмежувати значення слова w_i порівняно зі звичайними нейронними моделями, які використовують лише фіксоване вікно (тобто вони використовують лише часткову інформацію про тексти).

$$x_i = [c_l(W_i); e(W_i); c_r(W_i)] \quad (3)$$

Рекуррентна структура може отримати всі c_l при скануванні тексту вперед і c_r при зворотному скануванні тексту. Часова складність становить O(n). Після отримання представлення x_i слова w_i, ми застосовуємо лінійне перетворення разом з функцією активації tanh до x_i і відправляємо результат на наступний рівень.

$$y_i^{(2)} = \tanh\left(W^{(2)}x_i + b^{(2)}\right. \quad (4)$$

$y_{i(2)}$ - прихований смисловий вектор, в якому кожен семантичний фактор буде проаналізований для визначення найбільш корисного коефіцієнта для подання тексту.

Згорткова нейронна мережа в нашій моделі покликана представляти текст. З точки зору згорткових нейронних мереж, рекуррентна структура, яку ми згадували раніше, - це згортковий шар. Коли всі представлення слів обчислюються, ми застосовуємо шар максимального об'єднання.

$$y^{(3)} = max\, y_i^{(2)} \quad (5)$$

Функція max - це елементарна функція. K-й елемент $y_{(3)}$ - максимум в k-му елементах $y_{i(2)}$. Шар об'єднання перетворює тексти різної довжини у вектор фіксованої довжини. За допомогою шару об'єднання ми можемо захоплювати інформацію протягом усього тексту. Існують і інші типи об'єднання шарів, наприклад, середні шари об'єднання. Я не використовує тут середнє об'єднання, оскільки лише декілька слів та їх поєднання корисні для фіксації значення документа. Шар максимального об'єднання намагається знайти найважливіші латентні семантичні фактори в документі. Шар об'єднання використовує вихід вихідної структури, що повторюється, як вхід. Часова складність об'єднання шару становить O(n). Загальна модель являє собою каскад періодичної структури та шару максимального об'єднання, тому складність у часі нашої все ще O(n). Остання частина моделі - вихідний шар. Подібно до традиційних нейронних мереж, він визначається як:

$$y^{(4)} = W^{(4)} y^{(3)} + b^{(4)} \quad (6)$$

Наостанок, функція softmax застосовується до y(4). Вона може конвертувати вихідні числа у ймовірності.

$$p_i = \frac{\exp\,(y_i^{(4)})}{\sum_{k=1}^{n} \exp\,(y_k^{(4)})} \quad (7)$$

3 Експерименти та характеристики роботи системи

3.1 Принцип роботи

В результаті я отримав систему, яка на вхід приймає документ (мотиваційний лист абітурієнта). Далі проводиться попередня обробка текстових даних:

- Приведення всіх символів до нижнього регістру
- Токенізація
- Видалення стоп-слів
- Лематизація

Після того оброблений текст надходить до класифікатора на основі згорткових нейронних мереж. І на вихід отримуємо:

- Прізвище, ім’я, по-батькові абітурієнта
- 10 найбільш використаних слів
- Список фраз, які стосуються навчальних здобутків до університету.
- Декілька числових показників (значення від 0 до 1): грамотність, тональність (де 0 – негативна, 1 – позитивна)
- Загальна рекомендаційна оцінка якості мотиваційного листа абітурієнта.

3.2 Наступні кроки

Наступними кроками для вдосконалення робити системи, які не були оглянуті в межах цієї курсової робити буде додавання можливості подальшого тренування моделі. Точність результатів можна значно покращити збільшивши кількість тренувальних даних. Особливістю моделей

на основі глибинного навчання від традиційних алгоритмів машинного навчання є те що, їхня точність покращується постійно зі збільшенням даних, а ті в свою чергу доходять до певного порогу, за який неможливо вийти.

Для цього потрібно ввести механізми за яким людина, яка використовує систему, могла додатково виставляти оцінку якості мотиваційного листа. Вже на основі цих даних буде тренуватися модель, покращуючи свої результати в майбутньому.

Висновок

Метою цієї роботи була розробка системи автоматизованого аналізу мотиваційних листів для вступу у вищі навчальні заклади. У ході цієї робити були розглянуті сучасні методи робити таких технологій, проблеми обробки текстів природньою мовою. Сфери його застосування та принцип роботи.

Ця тема є досить актуальною в Україні оскільки у січні 2020 року в силу вступили зміни до закону про вищу освіту, який передбачає, що заклади вищої освіти прийматимуть мотиваційні листи, за які абітурієнт зможе отримати додаткові бали під час вступу.

В другому розділі було розглянуто процес розробки такої системи, необхідні кроки для досягнення точних результатів та алгоритм роботи класифікатора на основі згорткових нейронних мереж.

В третьому розділі описані принципи роботи такої системи та подальші можливі для вдосконалення її роботи.

Перспективно отриманий результат можу бути використаний у роботі приймальних комісій університетів.

Список використаної літератури

1. Bing Liu. Sentiment Analysis and Opinion Mining, Morgan & Claypool Publishers. – 2012.
2. Sentiment analysis: Capturing favorability using natural language processing: конф. Proceedings of the 2nd International Conference on Knowledge Capture (K-CAP 2003), Nasukawa, Yi. 23-25 Жовтня 2003.
3. Text Classification: [Електронний ресурс]. – Режим доступу: https://monkeylearn.com/text-classification/
4. Siwei Lai, Liheng Xu, Kang Liu, Jun Zhao. Recurrent Convolutional Neural Networks for Text Classification. – 2003