УДК 004.7

Бевз Д.М., студент І курсу СО «Магістр»
спеціальності 122 «Комп'ютерні науки»
Нескородєва Т.В., завідувач кафедри
інформаційних технологій

АНАЛІЗ МОДЕЛЕЙ ОБРОБКИ ТЕКСТІВ

Донецький Національний університет імені Василя Стуса, м. Вінниця

В останні роки широке впровадження інформаційних технологій призвело до збільшення обсягів текстової інформації, доступної в цифровому вигляді. Це створило потребу у розробці ефективних методів обробки текстів, які здатні автоматично розпізнавати, класифікувати та аналізувати цю інформацію. Аналіз текстів є важливим завданням в багатьох галузях, включаючи мовознавство, маркетинг, медіа, фінанси та інші.

1. Види та класифікації текстів.

Перед початком обробки текстів необхідно спочатку класифікувати їх за різними критеріями. Одним з таких критеріїв є тип тексту або його призначення. Розрізняють наступні види текстів:

- Інформативні тексти: новини, статті, наукові дослідження тощо;
- Літературні тексти: романи, вірші, оповідання;
- Технічні тексти: технічні описи, інструкції, технічні звіти;
- Рекламні тексти: оголошення, рекламні брошури, рекламні статті;
- Соціальні медіа-тексти: повідомлення в соціальних мережах, коментарі, пости.

Крім типу тексту, текстові дані можна класифікувати за тематикою, мовою, жанром, тощо. Це дозволяє виконувати більш специфічні завдання обробки та аналізу.

2. Етапи обробки текстів.

Обробка текстів передбачає кілька етапів, які можуть включати:

- Токенізація: розбиття тексту на окремі токени (слова, речення, символи) для подальшого аналізу;
- Лематизація і стемінг: перетворення слова до його базової форми для зниження розмірності і покращення збігу між словами.;
- Усунення стоп-слів: вилучення загальних слів (наприклад, "the", "is", "and"), які не мають вагомого значення для аналізу;
- Векторизація: перетворення тексту в числову форму, яка може бути використана алгоритмами машинного навчання. Це може включати TF-IDF, Bag-of-Words, Word2Vec та інші методи;
- Класифікація: призначення тексту до певної категорії або класу на основі його характеристик та використання моделей машинного навчання.;
- Аналіз сентименту: визначення емоційного відтінку тексту (позитивний, негативний, нейтральний);

- Екстракція іменованих сутностей: визначення та виділення іменованих сутностей, таких як особи, місця, організації, дати, тощо;
- Машинний переклад: автоматичний переклад тексту з однієї мови на іншу.

Аналіз текстів є важливим завданням у сфері обробки природної мови (Natural Language Processing, NLP). Обробка природних мов (англ. Natural Language Processing, NLP) створення систем з ознаками штучного інтелекту, які певним чином обробляють мовну інформацію з метою виконання певних задач [1]. До таких задач належать:

- чат-боти або формування відповідей на запитання користувача;
- визначення характеру емоційного забарвлення висловлювань;
- машинний переклад з однієї мови на іншу;
- розпізнавання мов;
- перевірка правопису;
- визначення частин мови в реченні і їх анотування;
- рерайт текстової інформації для створення веб-контенту [2].

Машинне навчання являє собою актуальну сферу наукового знання, яка інтенсивно розвивається та має дуже значні перспективи [3]. Розвиток нейромережевих моделей у сфері обробки текстів відкрив нові можливості для ефективного аналізу та розуміння текстової інформації. У даній статті ми розглянемо нейромережу Recurrent Neural Network (RNN) та її застосування в аналізі текстів. Детально проаналізуємо, що подається на вхід RNN та який має бути результат на виході.

Нейромережа Recurrent Neural Network (RNN)

RNN є типом нейромережі, спеціально розробленої для моделювання послідовностей даних, таких як текстові дані. Одна з особливостей RNN полягає у тому, що вона може зберігати попередні стани та використовувати їх для обробки наступних вхідних елементів послідовності.

На вхід RNN подається послідовність даних, наприклад, послідовність слів у тексті. Кожен елемент послідовності (слово) кодується у векторну форму, щоб його можна було обробити нейромережею. RNN проходить через кожен елемент послідовності по одному, кожен раз оновлюючи свій внутрішній стан залежно від поточного вхідного елементу та попереднього стану.

Результатом на виході RNN може бути різноманітна інформація, залежно від поставленої задачі. Наприклад:

Класифікація тексту: RNN може класифікувати текст у певні категорії або класи. Наприклад, у задачі аналізу електронних листів, RNN може визначати, чи є лист спамом чи не спамом.

Аналіз сентименту: RNN може визначати емоційний відтінок тексту, наприклад, позитивний, негативний або нейтральний. Це може бути корисним для аналізу соціальних медіа або відгуків користувачів.

Машинний переклад: RNN може використовуватись для машинного перекладу, де вона приймає на вхід текст у вихідній мові та генерує відповідний переклад у цільову мову.

Висновок.

Аналіз текстів є важливою задачею в сучасному інформаційному суспільстві. Розробка ефективних методів обробки текстів відіграє ключову роль у витягуванні корисної інформації з великого обсягу текстових даних. У даній статті ми розглянули види та класифікації текстів, а також основні етапи обробки текстів. Крім того, був наведений приклад використання нейромережі для обробки текстів. Опанування та вдосконалення методів обробки текстів допоможе в розвитку автоматичного аналізу текстів та покращенні різних застосувань, від інформаційного пошуку до машинного перекладу.

Список використаної літератури

1. *Natural Language Processing, NLP [Електронний ресурс] – Режим доступу до ресурсу: https://evergreens.com.ua/ua/articles/natural-language-processing.html .*
2. *Автоматична обробка текстів природною мовою та комп'ютерна лінгвістика : книга. посібник. / Є. Большакова, Е. Клишинський, Д. Ланде, А. Носков, О. Пескова та Є. Ягунова. – М. : МІЕМ, 2011. – 272 с*
3. *Уоссермен Ф. Нейрокомп'ютерна техніка / Ф. Уоссермен. – М. : Мир, 1992. – 238 с.*

УДК 004.01

Бойко У. В., Студент 1 курсу спеціальності 122 «Комп'ютерні науки» СО Магістр
Нескородєва Т. В., д.т.н., доцент кафедри інформаційних технологій

ІНТЕЛЕКТУАЛЬНИЙ АНАЛІЗ ДАТАСЕТУ «AUTO»

Донецький національний університет імені Василя Стуса, м. Вінниця

Вплив різних факторів на витрати палива автомобілів є актуальною та важливою проблемою у сучасному світі, де стала популярною енергоекономія та зменшення негативного впливу на довкілля.

В даній дослідженні було проведено аналіз набору даних Auto, що містить інформацію про різні характеристики автомобілів, з метою встановлення залежності між цими характеристиками та витратами палива.

Набір даних Auto складається з 392 спостережень за наступними 9 змінними:

- mpq – кількість пройдених миль на одному галоні палива;
- cylinders – кількість циліндрів у автомобіля;
- displacement – об'єм двигуна в кубічних дюймах;
- horsepower – потужність двигуна;
- weight – вага автомобіля;
- acceleration – час розгону від 0 до 60 миль/год в секундах;