УДК 004.855.5 DOI: 10.35803/1694-5298.2021.3.372-380

YOUNG-SANG CHOI, ABDIEVA S.
[1]KSUCTA n. a. N. Isanov, Bishkek, Kyrgyz Republic

ЙОНГ САНГ ЧОЙ, АБДИЕВА С.
[1]КГУСТА им. Н. Исанова, Бишкек, Кыргызская Республика
choi2018kz@gmail.com, a_sama09@mail.ru

A STUDY ON THE SENTIMENT ANALYSIS (POSITIVE, NEGATIVE) OF WORDS APPEARING IN KYRGYZ NEWS BY APPLYING THE DEEP LEARNING-BASED NLP (NATURAL LANGUAGE PROCESSING) TECHNIQUES FOR STUDENTS PRACTICE

ИССЛЕДОВАНИЕ АНАЛИЗА (ПОЛОЖИТЕЛЬНЫХ, ОТРИЦАТЕЛЬНЫХ) СЛОВ В КЫРГЫЗСКИХ НОВОСТЯХ С ПРИМЕНЕНИЕМ МЕТОДОВ НА ОСНОВЕ ГЛУБОКОГО ОБУЧЕНИЯ НЛП (ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА) ДЛЯ ПРАКТИКИ СТУДЕНТОВ

Бул изилдөө дүйнөнүн алдыңкы технологиясы болгон табигый тилди терең үйрөнүүгө негизделген сезимдерди талдоо талаасы боюнча теориялык, тактап айтканда маалыматтарды чогултуу жана алдын-ала иштетүү баскычы, токенизация этабы, Сезим сөздүгүн куруу этабы, сезимдерди талдоо аркылуу оң жана терс сөздөрдү алуу этабы, терең билим моделдин конфигурациясы, аткаруу этабы жана маалыматтарды визуалдаштыруу этабы сыяктуу негизги мазмундарды жана ага байланыштуу технологияларды киргизет.
Мындан тышкары, маалыматтарды чогултуу этабында аткарылган сүйлөөнү иштеп чыгуу технологиясы, STT (Speech to Text) жана TTS (Text to Speech) технологиялары киргизилет.
Бул изилдөөдө, ар кандай ачык булактардын (китепканалардын) жардамы менен, питон тилинин базасын табигый тилди иштетүү чөйрөсүндө колдонулган 'keras' 2.0 версиясы колдонулган программа жазылган. Бул изилдөө табигый тилди иштетүүнү терең үйрөнүп жаткан студенттерге жардам берүү максатында жүргүзүлдү.
***Өзөк сөздөр**: сезимдерди талдоо жана сөздүк, маалыматтарды алдын-ала иштеп чыгуу, терең үйрөнүү, визуалдаштыруу.*

Это исследование является теоретическим в области анализа отношений обработки естественного языка на основе глубокого обучения, которая является передовой мировой технологией, а именно этапом сбора и предварительной обработки данных, этапом токенизации, этапом построения словаря отношений, этапом извлечения положительных и отрицательных слов с помощью анализа отношений. Глубокое обучение знакомит с основным содержанием и связанными технологиями, такими как конфигурация модели, этап выполнения и этап визуализации данных.
Кроме того, будет представлена технология обработки речи, выполняемая на этапе сбора данных, технология STT (преобразование речи в текст) и TTS (преобразование текста в речь).
В этом исследовании программа была написана с использованием различных открытых источников (библиотек), включая версию 'keras' 2.0, используемую в области обработки естественного языка для глубокого обучения на базе языка Python. Это исследование выполнено, чтобы помочь студентам, изучающим глубокую обработку естественного языка.
***Ключевые слова:** анализ отношений и словарь, предварительная обработка данных, глубокое обучение, визуализация.*

This study is theoretical on the sentiment analysis field of deep learning-based natural language processing, which is the world's advanced technology, namely data collection and preprocessing stage,

ABBYY PDF Transformer 2.0 Click here to buy www.ABBYY.com

tokenizing stage, Sentiment Dictionary construction stage, positive and negative word extraction stage through sentiment analysis, deep learning introduces major contents and related technologies such as model configuration, execution stage, and data visualization stage.
In addition, speech processing technology performed in the data collection stage, STT (Speech to Text) and TTS (Text to Speech) technology will be introduced.
In this study, a program was written using various open sources (libraries) including ' keras' 2.0 version used in the deep learning natural language processing field of the python language base. This study is executed to help students who are studying deep learning natural language processing.
Key words: *sentiment analysis and dictionary, data preprocessing, deep learning, visualize.*

1. Introduction

Sentiment Analysis in natural language processing is an analysis method that quantifies sentiments as positive or negative based on sentiment polarity, such as subjective sentiments, attitudes, and tendencies in unstructured texts, and is widely used in the field of natural language processing.
Sentiment Analysis has been mainly performed in two approaches. The first is a lexical approach that performs sentiment analysis using a sentiment dictionary, and the second is a machine learning approach that classifies sentiments by applying machine learning techniques.
The lexical approach is a technique that is based on a vocabulary in which the criteria are defined in advance when performing sentiment analysis of sentences, paragraphs, or documents made up of words. Therefore, before performing sentiment analysis with a vocabulary-based approach, a sentiment dictionary is defined, and the sentiment score of the text data, which is the subject of sentiment analysis, is calculated and classified as positive or negative. In this method, sentiment analysis can be performed relatively easily after the sentiment dictionary is built, but the sentiment dictionary must be built in advance.
The machine learning-based approach refers to a model that extracts features so that a machine learning methodology can be used for text data, and trains a machine learning model based on the features of the extracted text data to perform sentiment analysis.
Machine learning sentiment analysis is performed through various techniques of supervised, unsupervised, and semi-supervised learning. Among them, the representative techniques of supervised learning used in sentiment analysis include Naive Bayesian classification, SVM (support vector machine), decision tree, and artificial neural network.
Recently, research on sentiment analysis using deep learning-based DNN(deep neural network) is being conducted. In particular, positive and negative binary classification sentiment analysis through CNN (Convolution Neural Network) was performed, and meaningful results were achieved. In addition, sentiment analysis was attempted using the LSTM (Long short-term memory) model, which effectively solved the long-term memory problem of RNN (Recurrent Neural Network) in a cell structure method.
In addition, in natural language processing for deep learning, there is an embedding process that converts natural language into a vector value composed of numbers so that the computer can understand the natural language spoken by humans.
According to the embedding method, each word or sentence is expressed in n dimensions. If a word or sentence is expressed as a unique vector in this way, similarity can be calculated through the vector operation process for each word or sentence, and a computer can grasp semantic and grammatical relationships just as humans infer. Before 2017, embedding techniques were mainly word-level, and the representative word embedding is Word2vec.
Word2vec is a method of receiving corpus input and vectorization of words in the corpus. The neural network algorithm used for learning word2vec is a natural language processing model that started from Harris's distribution hypothesis. word2vec's learning methods include CBOW (Continuous Bag of Words) and skip-gram model. The CBOW method predicts a specific word using the context composed of words located around the specific word, and the skip-gram predicts a word that can be located around the specific word based on the specific word.

In this study, from recording the news of Kyrgyzstan to collecting and refining data, building a sentiment dictionary, constructing a deep learning model, and processing data in detail, it can be used for deep learning natural language processing learning for students studying natural language. The overall processing process, the libraries (open sources) used in each step, and the results of the Python program are described.

2. Data collection and preprocessing

2.1 Data collection through STT (Speech to Text) technology

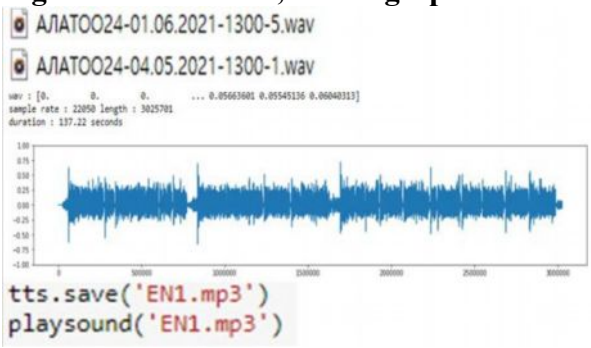
Data collection of natural language processing uses web scraping techniques from the Internet or pre-made text, but in this study we used news broadcast on YouTube 'Ала-Тоо 24-Новости Кыргызстана on YouTube 'Ала-Тоо 24-Новости Кыргызстана ' . These are May 4, 2021 13:00 news, May 22 13:00 news, May 24 13:00 news, May 25 13:00 news, May 27 13:00 news, 6 It is 13:00 news on the 1st of the month. The news was recorded in 'wav' format and converted into text using various STT techniques.

As libraries of the STT technology ‘librosa’ and ‘listdir’ were used to improve the sound quality so that the recorded voice could be heard well, and this was visualized in a graph.

For visualization of audio sound, FFT (Fast Fourier Transform) was used. This is to decompose the time-domain waveform through FFT (Fast Fourier Transform) and convert it into frequency-domain to grasp the degree of frequency forming the original sound data and visualize it.

And it was converted voice to text using ‘speech_recognition’. For reference, considering the performance of ‘speech_recognition’, the recording file size was limited to 2 minutes and 30 seconds. After that, Russian news recorded using Google APIs 'googletrans', 'gTTS', 'google playsound', and 'ipd.Audio' was converted into English or other languages and used to listen to audio while reading the translated text.

Figure 1. Audio files, Audio graph and translated English files



2.2 Data preprocessing and extraction of positive and negative words

In data preprocessing, 'speech_recognition' was used to convert Russian news into Russian text.

Figure 2. Russian text by applying STT technology

```
[ Converting audio transcripts into text ]
завершилась Рабочая поездка президента садыра жапарова в баткенской области накануне глава государства встретился с жителями се
ла Кулунда илекского района и села автобус Воткинского района во время встречи глава государства призвал бакинцев не поддаватьс
я на различные провокационные инсинуации относительно приграничных вопросов и не верить недостоверной информации и еще раз отме
тил что все пострадавшие дома будут безвозмездно восстановлены строительные работы предполагается завершить до июня семьям поги
бших выплатят компенсации по миллиону сумов а получившим ранения от 50 до 100 тысяч сумов он подчеркнул что с получением особог
о статуса баткенской области будет предоставлен ряд льгота дети пострадавших в пограничном конфликте семьи пройдут реабилитацию
на озере иссык-куль также президент добавил что продолжаются мероприятия по улучшению условий труда и повышению заработной плат
ы сотрудникам силовых структур в том числе пограничником По его словам в все время все силы будут брошены на полное решение при
граничных проблем Я буду направлять особые усилия для поддержки баткенской области Ну не позволим кому-либо претендовать на кыр
гызские земли особо отметил глава государства в ходе беседы с президентом местные жители интересовались вопросами дальнейшей по
ддержки со стороны государства трансформации земель и другими
```

For Russian sentences converted to text, removing stop words and word tokenizing were performed using Tokenizer and ‘text_to_word_sequence’ of ‘tensorflow.keras’, a deep learning library.

Figure 3. Tokenized Russian

```
tokenize:
 ['завершилась', 'рабочая', 'поездка', 'президента', 'садыра', 'жапарова', 'в', 'баткенской', 'области', 'накануне', 'глава',
'государства', 'встретился', 'с', 'жителями', 'села', 'кулунда', 'илекского', 'района', 'и', 'села', 'автобус', 'воткинског
о', 'района', 'во', 'время', 'встречи', 'глава', 'государства', 'призвал', 'бакинцев', 'не', 'поддаваться', 'на', 'различны
е', 'провокационные', 'инсинуации', 'относительно', 'приграничных', 'вопросов', 'и', 'не', 'верить', 'недостоверной', 'информ
ации', 'и', 'еще', 'раз', 'отметил', 'что', 'все', 'пострадавшие', 'дома', 'будут', 'безвозмездно', 'восстановлены', 'строите
льные', 'работы', 'предполагается', 'завершить', 'до', 'июня', 'семьям', 'погибших', 'выплатят', 'компенсации', 'по', 'миллио
ну', 'сумов', 'а', 'получившим', 'ранения', 'от', '50', 'до', '100', 'тысяч', 'сумов', 'он', 'подчеркнул', 'что', 'с', 'получ
ением', 'особого', 'статуса', 'баткенской', 'области', 'будет', 'предоставлен', 'ряд', 'льгота', 'дети', 'пострадавших', 'в',
'пограничном', 'конфликте', 'семьи', 'пройдут', 'реабилитацию', 'на', 'озере', 'иссык', 'куль', 'также', 'президент', 'добави
л', 'что', 'продолжаются', 'мероприятия', 'по', 'улучшению', 'условий', 'труда', 'и', 'повышению', 'заработной', 'платы', 'со
трудникам', 'силовых', 'структур', 'в', 'том', 'числе', 'пограничником', 'по', 'его', 'словам', 'в', 'все', 'время', 'все',
'силы', 'будут', 'брошены', 'на', 'полное', 'решение', 'приграничных', 'проблем', 'я', 'буду', 'направлять', 'особые', 'усили
я', 'для', 'поддержки', 'баткенской', 'области', 'ну', 'не', 'позволим', 'кому', 'либо', 'претендовать', 'на', 'кыргызские',
'земли', 'особо', 'отметил', 'глава', 'государства', 'в', 'ходе', 'беседы', 'с', 'президентом', 'местные', 'жители', 'интерес
овались', 'вопросами', 'дальнейшей', 'поддержки', 'со', 'стороны', 'государства', 'трансформации', 'земель', 'и', 'другими']
```

Then, Russian the number of counters of words and frequency of words were extracted using 'from collections import Counter' and 'from nltk.tokenize import TreebankWordTokenizer'.

Figure 4. Russian frequency of words

```
word conut:
 OrderedDict([('завершилась', 1), ('рабочая', 1), ('поездка', 1), ('президента', 1), ('садыра', 1), ('жапарова', 1), ('в',
5), ('баткенской', 3), ('области', 3), ('накануне', 1), ('глава', 3), ('государства', 4), ('встретился', 1), ('с', 3), ('жите
лями', 1), ('села', 2), ('кулунда', 1), ('илекского', 1), ('района', 2), ('и', 5), ('автобус', 1), ('воткинского', 1), ('во',
1), ('время', 2), ('встречи', 1), ('призвал', 1), ('бакинцев', 1), ('не', 3), ('поддаваться', 1), ('на', 4), ('различные',
1), ('провокационные', 1), ('инсинуации', 1), ('относительно', 1), ('приграничных', 2), ('вопросов', 1), ('верить', 1), ('нед
остоверной', 1), ('информации', 1), ('еще', 1), ('раз', 1), ('отметил', 2), ('что', 3), ('все', 3), ('пострадавшие', 1), ('до
ма', 1), ('будут', 2), ('безвозмездно', 1), ('восстановлены', 1), ('строительные', 1), ('работы', 1), ('предполагается', 1),
('завершить', 1), ('до', 2), ('июня', 1), ('семьям', 1), ('погибших', 1), ('выплатят', 1), ('компенсации', 1), ('по', 3), ('м
иллиону', 1), ('сумов', 2), ('а', 1), ('получившим', 1), ('ранения', 1), ('от', 1), ('50', 1), ('100', 1), ('тысяч', 1), ('о
н', 1), ('подчеркнул', 1), ('получением', 1), ('особого', 1), ('статуса', 1), ('будет', 1), ('предоставлен', 1), ('ряд', 1),
('льгота', 1), ('дети', 1), ('пострадавших', 1), ('пограничном', 1), ('конфликте', 1), ('семьи', 1), ('пройдут', 1), ('реабил
итацию', 1), ('озере', 1), ('иссык', 1), ('куль', 1), ('также', 1), ('президент', 1), ('добавил', 1), ('продолжаются', 1),
('мероприятия', 1), ('улучшению', 1), ('условий', 1), ('труда', 1), ('повышению', 1), ('заработной', 1), ('платы', 1), ('сотр
удникам', 1), ('силовых', 1), ('структур', 1), ('том', 1), ('числе', 1), ('пограничником', 1), ('его', 1), ('словам', 1), ('с
```

To extract positive and negative words using 'TextBlob', Russian text was converted into English text. The reason is that English has been studied grammatically the most in the deep learning natural language processing field, and there are many testable libraries.

Figure 5. Translated English text

```
[ Translated text ]
meanwhile, 83,685 people were vaccinated against coronavirus in the country, of which 30 5585 were vaccinated with a 2 dose,
Anya Rakhmatova also said at the briefing. For all questions regarding vaccination, you can contact the immunization departme
nt of the FMC, the Republican Center for Immunoprophylaxis and the Bishkek City Center for Immunization with the support of t
he national Red Crescent Societies have created a call center, it works daily from 8:00 am to 6:00 pm, it is planned to devel
op ecological tourism in Kyrgyzstan, the Deputy Director of the Tourism Department under the Ministry of Economy and Finance
said. a joint action plan was approved, one of the significant points of which will be the development of ecotourism, she sai
d money matova also noted that today the department is developing a plan for the restoration of the tourism sector after a pa
ndemic for 5 years. ecotourism is nature oriented tourism including the program Amma of environmental education and Enlighten
ment and carried out in accordance with the principles and environmental sustainability State Registration Service began issu
ing urgent general civil passports of the model of 2020 start accepting applications for issuing biometric passports From yes
terday, citizens can urgently issue and receive as soon as possible an international passport of Kyrgyzstan as soon as possib
le Regime a general civil passport of the sample of 20 years can be obtained in 3 hours 24 days and 8 days, the cost of servi
ces varies from 2300 soums to 6475 taking into account the bank's commission, we note the new generation passport complies wi
th the standards of the international civil aviation organization, it is protected by more than 30 modern security elements w
```

For the translated English text, removing stop words and word tokenizing were performed using the tokenizer and 'text_to_word_sequence' of 'tensorflow.keras'.

Figure 6. Tokenized English

```
 Translated text_to_word_sequence:
 ['meanwhile', '83', '685', 'people', 'were', 'vaccinated', 'against', 'coronavirus', 'in', 'the', 'country', 'of', 'which',
'30', '5585', 'were', 'vaccinated', 'with', 'a', '2', 'dose', 'anya', 'rakhmatova', 'also', 'said', 'at', 'the', 'briefing',
'for', 'all', 'questions', 'regarding', 'vaccination', 'you', 'can', 'contact', 'the', 'immunization', 'department', 'of', 't
he', 'fmc', 'the', 'republican', 'center', 'for', 'immunoprophylaxis', 'and', 'the', 'bishkek', 'city', 'center', 'for', 'imm
unization', 'with', 'the', 'support', 'of', 'the', 'national', 'red', 'crescent', 'societies', 'have', 'created', 'a', 'cal
l', 'center', 'it', 'works', 'daily', 'from', '8', '00', 'am', 'to', '6', '00', 'pm', 'it', 'is', 'planned', 'to', 'develop',
'ecological', 'tourism', 'in', 'kyrgyzstan', 'the', 'deputy', 'director', 'of', 'the', 'tourism', 'department', 'under', 'th
e', 'ministry', 'of', 'economy', 'and', 'finance', 'said', 'a', 'joint', 'action', 'plan', 'was', 'approved', 'one', 'of', 't
he', 'significant', 'points', 'of', 'which', 'will', 'be', 'the', 'development', 'of', 'ecotourism', 'she', 'said', 'money',
'matova', 'also', 'noted', 'that', 'today', 'the', 'department', 'is', 'developing', 'a', 'plan', 'for', 'the', 'restoratio
n', 'of', 'the', 'tourism', 'sector', 'after', 'a', 'pandemic', 'for', '5', 'years', 'ecotourism', 'is', 'nature', 'oriente
d', 'tourism', 'including', 'the', 'program', 'amma', 'of', 'environmental', 'education', 'and', 'enlightenment', 'and', 'car
ried', 'out', 'in', 'accordance', 'with', 'the', 'principles', 'and', 'environmental', 'sustainability', 'state', 'registrati
```

Then, by using 'collections-Counter' and 'nltk.tokenize-TreebankWordTokenizer', the number of counters of words and frequency of words were extracted.

A 'collections-Counter' is a dictionary subclass for counting hashable objects. It is a collection where elements are stored as dictionary keys and their counts are stored as dictionary values.

'word_tokenize' and 'WordPunctTokenizer' are libraries of NLTK (Natural Language Toolkit) that tokenize corpus.

Figure 7. Frequency of words

```
[ Translated words Count with stopwords ]
english word Counter({',': 12, 'tourism': 3, 'civil': 3, 'passport': 3, 'vaccinated': 2, '30': 2, 'also': 2, 'Kyrgyzstan': 2,
'plan': 2, 'ecotourism': 2, 'environmental': 2, 'issuing': 2, 'general': 2, 'passports': 2, 'soon': 2, 'possible': 2, 'intern
ational': 2, 'days': 2, 'elements': 2, 'meanwhile': 1, '83,685': 1, 'people': 1, 'coronavirus': 1, 'country': 1, '5585': 1,
'2': 1, 'dose': 1, 'Anya': 1, 'Rakhmatova': 1, 'informed': 1, 'briefing.': 1, 'Red': 1, 'Crescent': 1, 'Societies': 1, 'creat
ed': 1, 'call': 1, 'center': 1, 'works': 1, 'daily': 1, '8:00': 1, '6:00': 1, 'pm': 1, 'planned': 1, 'develop': 1, 'ecologica
l': 1, 'Deputy': 1, 'Director': 1, 'Tourism': 1, 'Department': 1, 'Ministry': 1, 'Economy': 1, 'Finance': 1, 'said.': 1, 'joi
nt': 1, 'action': 1, 'approved': 1, 'one': 1, 'significant': 1, 'points': 1, 'development': 1, 'said': 1, 'money': 1, 'matov
a': 1, 'noted': 1, 'today': 1, 'department': 1, 'developing': 1, 'restoration': 1, 'sector': 1, 'pandemic': 1, '5': 1, 'year
s.': 1, 'nature': 1, 'oriented': 1, 'including': 1, 'program': 1, 'Amma': 1, 'education': 1, 'Enlightenment': 1, 'carried':
1, 'accordance': 1, 'principles': 1, 'sustainability': 1, 'State': 1, 'Registration': 1, 'Service': 1, 'began': 1, 'urgent':
1, 'model': 1, '2020': 1, 'start': 1, 'accepting': 1, 'applications': 1, 'biometric': 1, 'From': 1, 'yesterday': 1, 'citizen
s': 1, 'urgently': 1, 'issue': 1, 'receive': 1, 'Regime': 1, 'sample': 1, '20': 1, 'years': 1, 'obtained': 1, '3': 1, 'hour
s': 1, '24': 1, '8': 1, 'cost': 1, 'services': 1, 'varies': 1, '2300': 1, 'soums': 1, '6475': 1, 'taking': 1, 'account': 1,
'bank': 1, "'s": 1, 'commission': 1, 'note': 1, 'new': 1, 'generation': 1, 'complies': 1, 'standards': 1, 'aviation': 1, 'org
anization': 1, 'protected': 1, 'modern': 1, 'security': 1, 'use': 1, 'active': 1, 'cover': 1, 'electronic': 1, 'chip': 1, 'pe
```

Analytical sentiment can be divided into seven categories. The second category is positive-negative, the third category is positive-neutral-negative, the fourth category is neutral-happiness-sad-anger, and the fifth category is nerve-happiness-surprise (surpurise + gear). , sadness, anger (hate + anger), the 6th category is neutral-happiness-surprise-fear-sad-anger (hate + anger) and the 7th category is neutral-happiness-surprise-fear-sadness-hate-anger. In this study, positive and negative words of the second category were extracted using 'TextBlob', the extracted word names and number of words were calculated, and 'polarity' and 'subjectivity' were also calculated. And the TTR (Type-token ratio) was calculated using 'from lexical_diversity import lex_div as ld'.

'TextBlob' is a friendly front-end to the pattern and NLTK(Natural Language Toolkit) libraries. 'TextBlob' utilizes native Python objects and syntax to facilitate smooth operation. The Quick-start example shows how to simply process the text to be processed as a string, and NLP (Natural Language Processing) methods such as tagging a part of speech can be used in a string object.

Lexical diversity is one aspect of 'lexical richness' and refers to the ratio of different unique word stems (types) to the total number of words (tokens). The term is used in applied linguistics and is quantitatively calculated using numerous measures including TTR (Text-type ratio), and the measure of textual lexical diversity (MTLD).

TTR is a value obtained by dividing the total number of types by the total number of tokens. TTR is commonly used as a measure of lexical diversity.

Polarity is float which lies in the range of [-1,1] where 1 means positive statement and -1 means a negative statement. Subjective sentences generally refer to personal opinion, sentiment or judgment whereas objective refers to factual information. Subjectivity is also a float which lies in the range of [0,1].

The subjectivity analysis method is just a method of classifying a given text as objective or not.

Figure 8. positive and negative words

```
[ Words analyzed for sentiment ]
Positive_feebacks Count : 1
['first']
Negative_feedback Count : 64
```

Figure 9. Polarity and subjectivity

```
[ polarity and subjectivity of sentiment ]
polarity  : 0.13636363636363635
subjectivity  : 0.5
```

Figure 10. TTR (Type-token ratio)

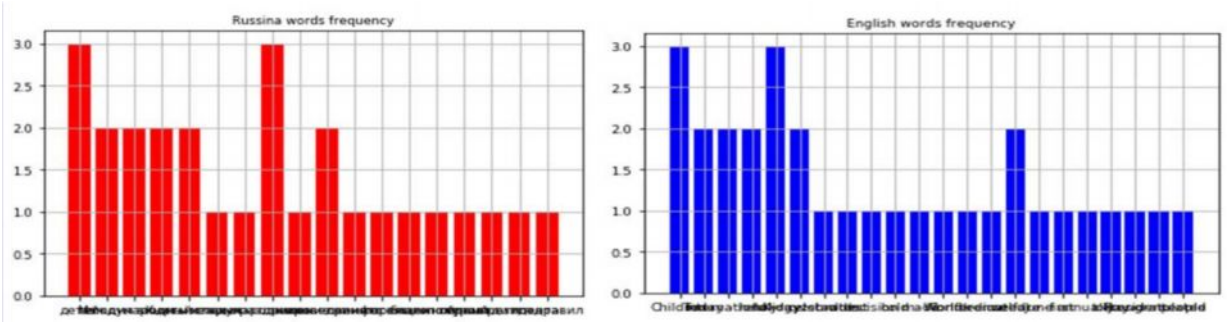
```
[ TTR ]
0.8333333333333334
```

Afterwards, Russian and English TF-IDFs were extracted using WordCloud and matplotlib. TF-IDF(The Term Frequency-Inverse Document Frequency) is a method of weighting the importance of each word in the DTM using the frequency of the word and the frequency of the inverse document (a specific expression is taken for the frequency of the document). The method of use is to first create a DTM, and then assign a TF-IDF weight. TF-IDF is mainly used for finding the similarity of documents, determining the importance of a search result in a search system, and for finding the importance of a specific word in a document.

Figure 11. TF-IDF in Russian and English

Using pandas and matplotlib, extract high frequency tokenized words and express them as histograms.

Figure 12. Histogram by word frequency



2.3 Create supermini-corpus

Based on the date order of the recorded files derived in the above process, a CSV (comma-separated values) file type supermini-corpus was created consisting of the derived number of tokens, number of words, number of positive words, number of negative words, TTR and sentiment values. Here, the sentiment value is composed of the number 0 or number 1, and the number 0 is a sentiment, whereas the number 1 is a sentence containing a positive sentiment. And using pandas, supermini-corpus of Excel base was converted into supermini-corpus of CSV type.

Table 1. Supermini-corpus

	Data	Tokens	Words	Positive	Negative	TTR	Class
0	20210405-1	428	305	5	300	0.71130	0
1	20210405-2	241	178	9	169	0.73780	1
2	20210405-3	531	386	13	373	0.72664	1
3	20210522-1	214	149	1	148	0.69355	0
4	20210522-2	234	201	10	191	0.85890	1

```
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Date      24 non-null     object
 1   Tokens    24 non-null     int64
 2   Words     24 non-null     int64
 3   Positive  24 non-null     int64
 4   Negative  24 non-null     int64
 5   TTR       24 non-null     float64
 6   Class     24 non-null     int64
dtypes: float64(1), int64(5), object(1)
```

2.4 Constructing supermini-Sentiment Dictionaries using word2vec and textblob.wordnet

If a word or sentence is expressed as a unique vector, similarity can be calculated through the vector operation process for each word or sentence, and a computer can grasp a semantic or grammatical

relationship just as a human infers. word2vec is a representative word embedding technique at the word level.

In this study, the steps to construct a supermini-sentiment dictionary using word2vec and 'textblob.wordnet' were set as follows. The first is to extract words from text data. The second step is to check the frequency of the extracted words. Third step, similar words of the extracted words are extracted. The fourth step is to extract the similarity between the extracted words. Here, the similarity between the two words increases as the sympathetic relationship with the same sentimental relationship between the two words increases, and the antipathy relationship with the opposite sensibility decreases. In the fifth step, a supermini-sentiment dictionary is created by integrating the sentiments of the inferred words.

The WordNet used above is a representative of thesaurus, classifies several words into a synonym group called 'synset' and provides a simple and general definition. It also records the various semantic relationships between lexicons.

Figure 13. Derived similar words and degree of similarity

```
word = Word("vaccination")
word.synsets
```

```
[Synset('inoculation.n.01'), Synset('vaccination.n.02')]
```

```
start01 = Synset('inoculation.n.01')
end01 = Synset('vaccination.n.02')
print('similarity :', start01.path_similarity(end01))
```

```
similarity : 0.07692307692307693
```

3. Deep Learning Model Configuration and Data Visualization

3.1 Deep Learning Model Construction

To build the deep learning model of this study, 'tensorflow.keras' and 'sklearn' were used to build the deep learning model. However, since the sample data is too small, it consists of 6 input layers, 8 hidden layers, and 1 output layer, 8 hidden layers, and 1 output layer, ReLU(rectified Liner Unit) function for input and hidden layers, and sigmoid for output layer. We set the function, Loss function uses binary_crossentropy, optimizer set to 'adam', metrics=['accuracy'] set, batch_size=5, epochs=50, and deep learning was run.

Here, the ReLU function is an activation function that returns 0 when the input value is negative, while returning the input value as it is when the input value is positive. It is a function that is used a lot recently because of its simple implementation. After that, when passing the result value from the last hidden layer to the output layer, the sigmoid function was used to pass the value to the output layer between 0 and 1. It is predicted that the closer the output layer's result value is to 1, it is a more positive. And it is predicted that the closer the output layer's result value is to 0, it is a more negative.

In the neural network learning process, the current state is expressed as a loss function index. Loss is a measure of the difference between the model predicted value and the actual data value. In this study, 'binary cross entropy', which is suitable for training a neural network that predicts one of two classes (positive or negative), is set as a loss function. And neural network learning is to find the optimal parameters (weight and bias) that lower the value of the loss function as much as possible based on this loss as a function index, and this process is called optimization. Since 'Adam' (Adaptive Moment Estimation) used to optimize is an optimization algorithm that mixes momentum and 'RMSprop', it is the most commonly used optimization algorithm in deep learning.

Figure 14. Deep learning run

```
Epoch 48/50
24/24 [==============================] - 0s 1ms/sample - loss: nan - accuracy: 0.6667
Epoch 49/50
24/24 [==============================] - 0s 1ms/sample - loss: nan - accuracy: 0.6667
Epoch 50/50
24/24 [==============================] - 0s 958us/sample - loss: nan - accuracy: 0.6667
```

As a result of learning with the training data, it showed an accuracy of 66.77

To increase accuracy, sklearn's 'train_test_split' library was used, and the ratio of training data was set to 70% randomly, the loss function was set to 'mean_squared_error', the optimizer was set to 'adam', metrics=['accuracy'], epochs=50, batch_size=5, and deep learning was run. And the learning rate and validation of the training data were not applied.

'mean_squared_error' defines the mean squared error between the predicted value and the actual value. The formula is very simple, the larger the difference, the clearer the value is due to the squaring operation. And squaring increases the cumulative value, whether the error is positive or negative.

Figure 15. Deep learning run with splitting training data and test data

```
Epoch 48/50
16/16 [==============================] - 0s 1ms/sample - loss: nan - accuracy: 0.8125
Epoch 49/50
16/16 [==============================] - 0s 1ms/sample - loss: nan - accuracy: 0.8125
Epoch 50/50
16/16 [==============================] - 0s 1ms/sample - loss: nan - accuracy: 0.8125
```

As a result of training by separating the training data and the test data, the accuracy was improved to 81,25.

3.2 Data Visualization

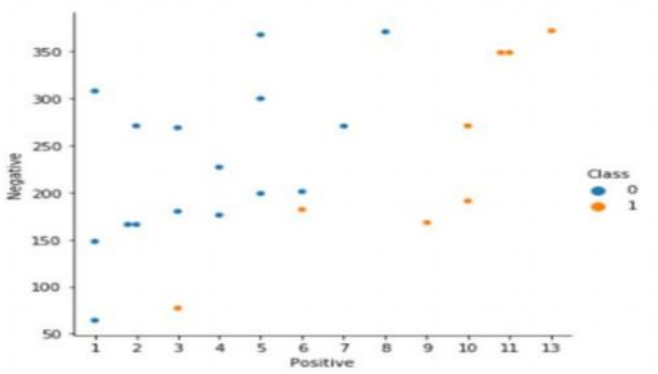
Visualization was performed by creating various graphs such as catplot, FacetGrid and heatmap using Matplotlib and Seaborn.

Matplotlib is a Python library that allows you to plot various data in many ways.

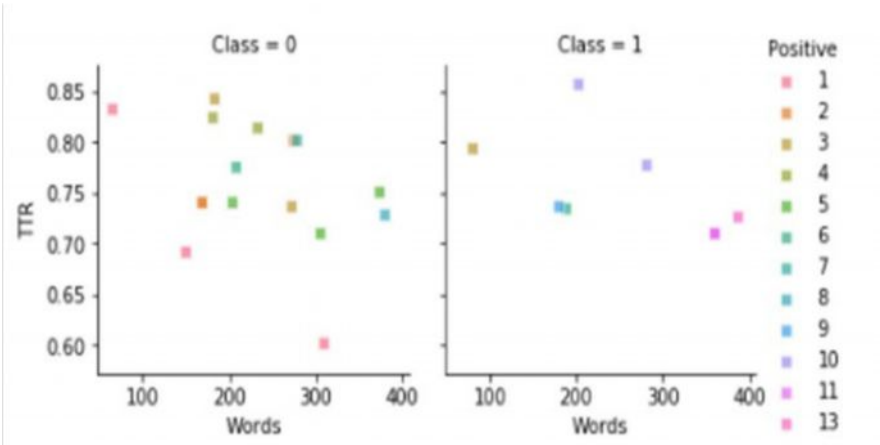
Seaborn is a visualization package based on Matplotlib that adds various color themes and charts for statistics. Basic visualization functions depend on the Matplotlib package and the statistics functions depend on the Statsmodels package.

Figure 16. Visualized graphs

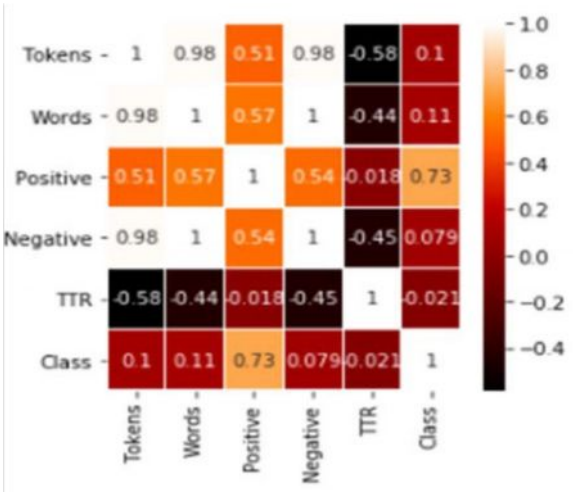
Using 'sns.catplot', x-axis is 'Positive' , y-axis is 'Negative'



Using 'sns.FacetGrid', x-axis is 'Words' , y-axis is 'TTR',



Using 'sns.heatmap', Show correlation of Tokens, Words, Positive, Negative, TTR, Class



4. Research Results

In this study, the whole process of sentiment processing in natural language processing is done using deep learning.

That is, data collection through STT(Speech to Text) technology, data pre-processing, positive and negative word extraction, supermini-corpus writing, constructing supermini-Sentiment Dictionary using word2vec and 'textblob.wordnet', deep learning model configuration, data visualization, etc. This was done by creating a Python program using open source(libraries).

As a result of the study, it was found that positive and negative messages can be analyzed through morphological analysis of words used in news.

And by classifying these words into positive and negative words, it was found that these words can be applied to sentiment analysis.

In the deep learning execution stage, 'tensorflow.keras' and 'sklearn' were used.

The first was to run without splitting the training data and the test data and the accuracy was calculated. The second was to run splitting into 70% training data and 30% test data and the accuracy was calculated.

The program written in this study used various libraries (or algorithms) such as ‘tensorflow.keras’ used in deep learning. If you learn usage of these and understand the ability to customize each library, you will improve your skills highly in the field of deep learning natural language processing.

References

1. Hobson Lane, Hannes Max Hapke, Coke Howard. Natural Language Processing in Action, Manning Publication, 2018.
2. Saito Goki, Deep Learning from Scratch2, O’Reilly Japan. 2018.
3. T.H. Jo, Deep Learning for everyone, Gilbut Publication, 2019.
4. E.S, Choi, A Thorough Introduction to Python for Data Analysis, Wikibooks, 2019.
5. SERGEY SMETANIN, The Applications of Sentiment Analysis for Russian Language Texts, IEEE date of publication June 15, 2020.
6. Neha Giaur and Neetu Sharma, Sentiment Analysis in Natural Language Processing, International Jouranal of Enginnering and Techniques, 2017.
7. Seo, Hye-Jin and Jeong-Ah Shin. Data processing and transformation in the sentiment analysis using a deep learning technique. Korean Journal of English Language and Linguistics Vol. 19. No 4. 2020
8. Heo, C. and S, On. A Novel Method for Constructing Sentiment Dictionaries using Word2vecand Label Propagation. Journal of Korean Institution of Next Generation Computing Vol13, No2. 2017.

9. K.N, Lee , How far has speech language processing technology come. Life in the New Korean Language, Vol. 27, No. 4. 2017