

전쟁뉴스 기사로부터 감정분석 및 위기 상황 탐지에 관한 연구

송윤아, 이유찬, 권현*

육군사관학교 AI·데이터과학과

songya030526@gmail.com, dldbcks64@gmail.com, *hkwon.cs@gmail.com(교신저자)

A Study on detecting emotions and crisis situations from war news articles

Song Yoon A, Lee Yu Chan, Hyun Kwon

Department of Artificial Intelligence and Data Science, Korea Military Academy

요약

본 논문은 전쟁뉴스 기사의 감정을 분석하고 위기 관련 단어 빈도를 통해 위기 지수를 계산하여 전쟁 발발 가능성을 예측하는 모델을 연구하였다. 뉴스 기사 본문의 감정분석을 위해 SpaCy 문장 분리 모델과 Bert-base-multilingual-uncased 모델을 사용하였다. 이후 위기 관련 단어 리스트를 통해 본문의 위기 단어 비율을 계산하여 위기 수준을 3가지로 분류하였다. 최종적으로 감정분석 및 위기 수준 결과를 통합하여 위기 지수를 계산하였다. 본 학습 모델은 우리나라의 안보 증진을 위한 전쟁 발발 사전 대응 시스템 구축에 기여할 수 있을 것으로 기대된다.

I. 서론

21세기 현대 전쟁은 4차 산업혁명으로 인한 빠른 정보의 흐름으로 기존 무력 충돌뿐만 아니라 심리전, 정보전, 여론전 등 다양한 양상을 나타내고 있다. 이러한 상황에서 정보를 통해 전쟁 발발 가능성을 조기에 예측하는 것은 효과적인 방어 태세 구축 및 즉각적인 대응을 가능하게 할 것이다. 2022년 발발한 러시아·우크라이나 전쟁이나 2023년 발발한 이스라엘·하마스 전쟁을 보면 전쟁의 전조가 수년간 언론 보도를 통해 드러났음에도 불구하고 이를 실질적으로 예측하지 못해 효과적인 대응에 실패하였다. 이에 본 논문에서는 전쟁 발발 이전 뉴스 기사들을 분석하여 전쟁 발발 가능성을 예측하는 모델을 개발하고자 하였다. 본 모델의 학습 과정은 다음과 같다. 우선, 전 세계에서 발발한 전쟁과 관련된 뉴스 기사들을 수집하고, 기사 내용을 문장 단위로 분리 후 불필요한 정보를 제거한다. 이후 각 문장의 감정을 분석하여 뉴스 기사의 감정 점수를 계산한다. 감정 점수와 전쟁 관련 위기 단어 빈도를 기반으로 최종적으로 위기 지수를 계산 및 학습하고 전쟁 발발 가능성을 예측하는 모델이다. 본 학습 모델은 단순히 과거 사건의 분석에 그치지 않고, 향후 전쟁 발발 대비 사전 대응 체계 시스템을 구축하는데 기여할 수 있을 것으로 기대된다.

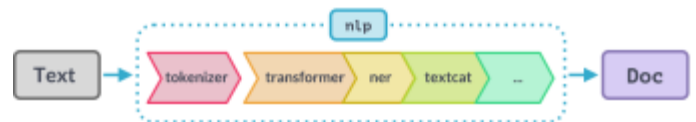
II. 감정분석

1. SpaCy

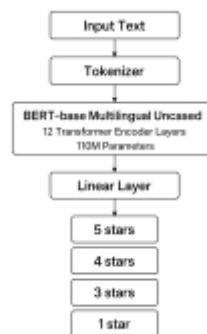
SpaCy는 산업용 자연어처리(Natural Language Processing, NLP)에 최적화된 고성능 오픈소스 파이썬 라이브러리이다.

주요 구성 요소는 문장을 토큰으로 분할하는 Tokenizer, 품사를 태깅하는 Tagger, 의존 구문을 분석하는 Parser, 명명 엔티티를 인식하는 NER, 표제어 추출을 하는 Lemmatizer, 문서를 분류하고 감정을 분석하는 Text Categorizer, 패턴에 기반하여 개체를 인식하고 규칙에 기반하여 탐지하는 Entity Ruler/Matcher이 있다.

SpaCy는 매우 빠른 속도와 다양한 언어 모델을 지원함으로써 뉴스 기사 분석과 감정분석, 정보 추출 시스템에 적용할 수 있다.



2. Bert-base-multilingual-uncased



본 모델은 감정 분석(Sentiment Classification)을 위해 사전 학습(pre-trained)된 BERT 기반 다국어 모델이다. 자연어 입력에 대해 1~5점의 감정 점수를 분류하는 목적으로 설계되었으며, 감정의 강도와 방향(긍정·부정)을 정량적으로 추정할 수 있는 장점이 있다.

Bert-base-multilingual-uncased 아키텍처를 기반으로 하며, Transformer 계열의 BERT 구조, 12개의 Transformer 인코더 레이어, 12의 self-attention head, 768 hidden dimension, 110M 파라미터 규모, Tokenizer로 되어있다. 해당 아키텍처는 100개 이상의 언어로 사전학습된 후, 감정 분류 작업을 위해 추가적인 파인튜닝이 이루어졌다.

모델의 출력은 5개의 클래스와 그 확률값이다. 5개의 클래스에서 1 star는 매우 부정적, 2 stars는 부정적, 3 stars는 중립, 4 stars는 긍정적, 5 stars는 매우 긍정적인 것을 나타낸다. 그리고 5개의 클래스에 대한 확률값도 같이 출력해서 softmax 함수를 이용해 계산한다.

본 연구에서 이 모델을 이용해 뉴스 기사에 대한 대중의 감정을 파악하였고, 위기 예측을 위한 지표의 감정적 강도 입력 요소로 활용하였다.

3. 감정분석 모델

우선, 전쟁 관련 뉴스 기사들의 URL을 수집하였다. 이후 기사 본문에서 40자 이상의 문단만을 추출하고 광고 등의 불필요한 정보는 제거하였다. 추출된 문단들은 문장 분리 라이브러리인 SpaCy를 통해 문장 단위로 분리하였다.

문장 단위로 감정분석을 하기 위해 Bert-base-multilingual-uncased 모델을 사용하였다. 본 모델은 각 문장에 부여된 점수를 기준으로 1점에서 2점은 Negative, 3점은 Neutral, 4점에서 5점은 Positive 총 세 가지 감정으로 분류한다. 이후 각 문장의 감정 분포 및 전체 평균 감정 점수를 계산하여 뉴스 기사의 전반적인 감정 경향성을 파악하고자 하였다.

III. 위기 단어 탐지

뉴스 기사를 통해 위기 상황을 정량적으로 분석하기 위하여 위기 단어 탐지를 수행하였다. 위기 단어 리스트는 기존 전쟁 관련 뉴스에서 자주 등장하는 단어들을 기반으로 구성되었고, 총 20개의 단어로 이루어졌다. 위기 단어 탐지를 위해 기사 본문을 소문자로 변환한 후 단어를 토큰화하였다. 전체 단어 수에서 위기 단어 비율을 계산하였고 이 비율에 따라 위기 수준을 분류하였다. 위기 수준은 총 3가지로 분류하였고 분류 기준은 다음과 같다. 위기 단어 비율이 1.5% 이상일 때 위기 상황(고위험 기사), 위기 단어 비율이 0.5% 이상, 1.5% 미만일 때 경계 상황(일부 위기 관련 기사), 위기 단어 비율이 0.5% 미만일 때 안정 상황(일반적 기사)으로 분류하였다.

IV. 위기 지수 계산

이전 단계에서 진행한 감정분석과 위기 단어 탐지에 기반하여 최종적으로 위기 지수를 계산하였다. 위기 지수는 문장별 감정 점수와 위기 단어 비율을 통합적으로 반영하기 위해 도입한 개념으로 위기 지수 수식은 다음과 같다.

$$\text{Crisis Index} = (5 - \bar{S}) \times R \times K$$

이 수식에서 S는 전체 기사 문장에 대한 평균 감정 점수로 1점에서 5점 사이의 값을 지닌다. R은 위기 단어 비율, K는 조정 계수로, 본 연구에서는 1로 설정하였다. 이 수식의 논리는 다음과 같다. 기사의 감정 점수가 5점에 가까울수록 기사의 감정이 긍정적이므로 (5-S) 값은 작아지고 반대로 1점에 가까울수록 기사의 감정은 부정적이므로 (5-S) 값은 커진다. 이 값에 위기 단어 비율 값인 R 값이 곱해져 최종적인 위기 지수를 계산한다. 위기 지수를 분류하는 기준은 총 3가지로 위기 지수가 5 이상이면 전쟁 경고 수준(심각한 위기), 위기 지수가 2 이상 5 미만이면 경계 상태(위험 신호 존재), 위기 지수가 2 미만이면 안정 상태(상대적으로 안전)으로 분류하였다.

V. 우리나라의 위기 지수 분석

본 모델을 통해 우리나라의 위기 지수 분석을 진행해 보았다. 본 연구의 최종 목적은 전쟁 발발 예측 및 사전 대응 시스템을 개발하는 것이므로 본 모델을 통해 우리나라의 현재 위기 지수를 분석해보아야 한다고 판단하였다. 우선, 북한의 위협 관련 기사 10개를 찾고 각각의 감정 점수 및 위기 지수를 계산하였다. 계산 결과 전체 평균 감정 점수는 2.42점, 전체 평균 위기 지수는 5.29라는 값이 나왔다. 즉, 전쟁 경고 수준(심각한 위기) 상태가 나왔다. 이 결과로부터 얻을 수 있는 사실이자 교훈은 우리나라는 현재 전쟁을 잠깐 멈춘 상태인 휴전국이라는 점이다. 우리나라는 북한의 지속적인 도발과 핵미사일 개발 등으로 인해 안보가 위태로운 상황에 있다. 우리나라는 현재의 평화에 안주하지 말고 북한의 안보 위협에 대비하여 확실한 대응 시스템을 구축해야 할 것이다.

VI. 결론

본 논문에서는 전쟁뉴스 기사로부터 감정 및 위기 상황 탐지에 관한 연구를 시행하였다. 본 연구에서는 전쟁뉴스 기사의 감정을 분석하고 기사 전문의 위기 단어를 탐지하였다. 감정분석 결과와 위기 단어 탐지 결과를 통합하여 최종적으로 위기 지수를 계산하고 모델을 통해 학습 및 평가를 진행하였다. 본 연구를 진행한 목적은 기존 각종 전쟁의 전조가 전쟁 발발 전 수년간 뉴스 기사를 통해 드러났음에도 전쟁을 예측하지 못하여 효과적인 대응에 실패하는 현실의 문제점을 극복하고자 하였다. 또한, 우리나라는 현재 전쟁을 잠깐 멈춘 상태인 휴전국이다. 우리나라는 북한의 지속적인 도발과 핵미사일 개발 등의 맞서 안보를 지켜야 한다. 이에, 전쟁 발발 예측을 위한 연구를 진행해 보았다. 본 연구의 학습 모델을 고도화된 기술과 더 많은 양의 데이터를 활용하여 학습시킨다면 뉴스 기사를 포함한 다양한 매체를 이용해 전쟁을 실질적으로 예측할 수 있을 것으로 기대된다. 전쟁 또한 인간의 심리에 의해 영향을 받으므로 전쟁의 정확한 발발 일자는 예측하기 어렵겠지만 본 학습 모델을 통해 전쟁 발발 예측 시스템을 발전시켜 매체의 감정분석 및 위기 관련 단어 탐지로 전쟁 발발 위험도가 높은지 낮은지 등을 정확하게 평가할 수 있게 된다면 국가의 안보를 더욱 증진시킬 수 있을 것이다.

참 고 문 헌

- [1] Hugging Face 모델 카드 : <https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment>
- [2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.
- [3] ExplosionAI (n.d.). *spaCy Documentation*. <https://spacy.io/usage>
- [4] Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing
- [5] 남영자. (2020). CNN을 활용한 방송 뉴스의 감정 분석. 한국정보통신학회논문지, 24(8), 1064-1070.
- [6] Alam, F., Joty, S., & Imran, M. (2018). Graph based semi-supervised learning with convolution neural networks to classify crisis related tweets. Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018), 556-559.