**Explaining the Deep Natural Language Processing by Mining Textual Interpretable Features**

[**https://arxiv.org/pdf/2106.06697.pdf**](https://arxiv.org/pdf/2106.06697.pdf)

1. **INTRODUCTION**

|  |
| --- |
| 다음 Table 1은 LSTM 모델에 의해 생성된 misleading prediction을 보여준다. 여기서 **두 문장 모두 Clean language이지만, prediction은 서로 완전히 다르다.**    여기서는 **T-EBANO (Text-Explaining BlAck-box mOdels)**라는 모델을 도입하며, 이것은 NLP에서의 black-box 신경망 모델이 생성한 결정을 이해할 수 있게 하는 프레임워크이다.   * T-EBANO는 서로 다른 interpretable feature에 적용되는 **perturbation process**를 통해 **prediction-local explanation**을 생성한다.   이 연구는 다음에 기여한다.   * **T-EBANO**라는 NLP task에 대한 XAI 방법론의 설계 및 개발   + Prediction-local 및 model-global한 설명을 모두 제공 * **Model-wise하고 해석 가능한 feature의 집합**을 통한, 입력되는 텍스트 데이터를 설명하는 효과적인 전략의 설계 * **양적 및 질적 설명**을 정의하고, 블랙박스 모델이 제공하는 local outcome에서 **각 feature set의 영향력**을 측정 * 혁신적인 model-global 설명 전략의 정의   + **Global Absolute Influence와 Global Relative Influence score**라는 새로운 metric을 이용한 Inter- 및 intra-class 영향력 분석 * 2개의 잘 알려진 블랙박스 신경망 구조인 **BERT와 LSTM에서의 철저한 실험적 평가** |

1. **LITERATURE REVIEW**

|  |  |
| --- | --- |
| Model-agnostic approach | **LIME, SHAP** 등   * 특정한 경우에는 model-agnostic technique가 **단일 단어의 prediction에 대한 impact**를 분석한다. |
| Domain-specific approach | Model reaction을 통해 prediction-local explanation을 생성 |
| Task-specific approach | |

1. **T-EBANO OVERVIEW**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| T-EBANO는 **NLP에 사용되는 black-box model의 inner functionality**를 설명한다.   * **Model-agnostic**과 **model-aware**의 서로 다른 building block을 포함한다.  |  |  | | --- | --- | | Model-agnostic | Part of speech, sentences | | Model-aware | Multi-layer word embeddings |     Classification task에 대해서 다음과 같이 수행한다.   1. 입력되는 문서는 **black-box 모델에 제공**된다. 2. T-EBANO는 **interpretable feature의 집합을 추출**한다. 3. NLP technique 또는 모델 자체의 숨겨진 지식에 대한 분석을 이용하여, 이들 **interpretable feature의 집합에 대한 perturbation**을 수행한다. 4. Interpretable feature에 대한 **perturbation은 모델의 출력에 서로 다른 방법으로 영향**을 끼칠 수 있는데, 그 방법은 다음과 같다.  |  |  | | --- | --- | | Case (a) | 분석 중인 class에 대한 **확률이 증가**한다.   * 분석된 feature는 **프로세스에 negative한 영향**을 미쳤다. | | Case (b) | Class에 대한 **확률이 감소**한다.   * Perturb된 feature는 **프로세스에 positive한 영향**을 미쳤다. | | Case (c) | Class에 대한 **확률이 거의 변하지 않는다.**   * 입력의 이 부분은 **예측 모델과 거의 관련이 없다.** |   **예측 과정에서의 perturbation 전후의 차이**의 중요성은 **nPIR index**를 이용하여 측정된다.   * nPIR는 perturbation 전략의 효과를 측정하기 위한 양적 방법이다. * 이것을 이용하여 T-EBANO는 **local explanation report**를 생성하고 그 결과를 대시보드를 통해 제공한다.   마지막으로, **생성된 local explanation을 aggregate**하여 **model-global explanation**을 생성한다.   * 블랙박스 모델의 결정에 영향을 미친 inter-, intra-class semantic 개념들을 highlight하여 보여준다. |

1. **INTERPRETABLE FEATURES**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| T-EBANO는 **word (tokens)와 sentence granularity level을 모두 고려**하여 interpretable feature의 집합을 추출한다. 또한, word의 appearance가 때로는 매우 중요하기 때문에 **입력된 텍스트상의 추출된 feature의 위치를 기억**한다.  T-EBANO는 다음의 **3가지 서로 다른 종류의 interpretable feature extraction technique**를 포함한다.   |  |  | | --- | --- | | Multi-layer Word Embedding (MLWE) | **모델이 학습한 inner knowledge를 이용**하는 것이므로 가장 강력한 전략이다.   * 모델의 inner knowledge에 접근하기 위해, **model의 inner detail**을 먼저 알아야 한다. * 이 방법은 서로 다른 딥러닝 구조 및 이들의 hidden layer에 쉽게 적용 가능하다. | | Part-of-Speech (PoS) | **어떤 part-of-speech에 속하는지**를 이용하여 단어의 semantic meaning을 탐색한다.   * **Input text가 tokenize**된다. (**token** itself, the **position** in the text, its **pos-tag**) * Token은 품사에 따라 correlated group으로 나누어진다. (adjectives, nouns, verbs, adverbs and others) | | Sentence-based feature extraction | **각 문장을 서로 독립적으로 고려**하여 모델의 결정에 대한 영향력을 측정한다.   * 모델이 **문장의 완전한 의미를 이해하고 이것을 통해 결과를 출력**하는지를 verify하는 전략이다. * Sentence feature extraction은 input text를 **문장에서의 position과 문장 그 자체**로 characterize한다. | |

1. **MULTI-LAYER WORD EMBEDDING (MLWE) FEATURE EXTRACTION**

|  |
| --- |
| Deep Neural Network는 학습 데이터로부터 복잡한 수학적 모델을 추출하는데, 이전에 확인하지 못한 unseen data에 대한 prediction process에서는 모든 레이어가 출력에 기여한다.   * 따라서 믿을 만한 설명을 위해서는 **모델의 모든 hidden layer로부터 모든 지식을 마이닝해야 한다.** * T-EBANO는 **Multi-layer Word Embedding (MLWE) feature extraction**을 통해서 이것을 수행한다. * T-EBANO는 **여러 층의 hidden layer의 출력을 분석**하고 그것의 **수학적 표현**을 네트워크의 서로 다른 레벨에서 추출한다. |

**4. MULTI-LAYER WORD EMBEDDING (MLWE) FEATURE EXTRACTION (Cont.)**

|  |
| --- |
| **T-EBANO MLWE feature extraction** 프로세스는 다음 그림과 같다.     1. **Input document**가 블랙박스 모델에 주어진다. 2. 서로 다른 layer에 있는, 서로 다른 word에 대한 **수학적으로 임베딩된 representation을 포함한 tensor**를 추출한다. 3. 각 레이어의 **intermediate embedding이 aggregate**된다. (합계 또는 평균) 4. Aggregation step의 출력은 input document에 대한 Multi-layer Word Embedding 표현이다. 5. **서로 비슷한 MLWE를 갖는 word들은 model이 판단하기에 서로 관련성이 높다.** 또한 함께 **group이 되어 있으면 그것들은 key input concept**를 표현하고, 이것들은 현재의 예측에 강력하게 영향을 미치는 것으로 보인다. 6. MLWE가 extract된 후에는 그것들은 **unsupervised clustering**을 이용하여 분석되고, 서로 비슷한 행동을 공유하는 **연관된 word의 set을 identify**한다.    1. 이러한 unsupervised analysis의 목적은 **모델의 출력에 가장 큰 영향을 미치는 input word의 가장 작은 그룹을 identify**하는 것이다.   MLWE feature extraction, 특히 여러 층에서의 aggregated word embedding에 대한 extraction은 **신경망의 구조에 따라 서로 다른 방법**으로 이루어진다.  T-EBANO는 K-Means clustering을 적용하여 **개의 group을 identify**하며, 여기서 **는 cluster의 최대 개수**로 다음과 같다.  이 방법의 목적은 **best explanation**만을 사용자에게 제공하는 것이므로, **partitioning된 cluster의 개수 K**의 값에 대해 **normalized Perturbation Influence Relation (nPIR) index**를 계산한다.  이 프로세스의 최종 출력은 **most informative local explanation**이다. |

1. **INTERPRETABLE FEATURE PERTURBATION**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Interpretable feature set을 추출한 다음에는 **perturbation phase를 실행**한다.   * 이때 **noise를 추가**하는데, 이것은 서로 다른 state-of-the-art 접근 방식에 의한 잘 알려진 테크닉이다. * Textual data에서는 perturbation은 **feature removal 또는 feature substitution**에 의해 동작한다.  |  |  | | --- | --- | | Feature removal | T-EBANO가 제공하는 feature removal 접근 방법에서 **모든 interpretable feature들은 input text에서 반복적으로 제거**되며, 새로운 perturbed variation을 생성한다.   * 이들 **perturbed variation은 모델로 피드백**되며, 이들에 대한 분석은 수집 및 T-EBANO에 의해 분석된다. * 이를 통해 **local explanation report를 생성**한다. | | Feature substitution | Removal perturbation이 제거된 word와 관련된 concept를 없앨 수 있지만, substitution perturbation은 이와 관련 있는 **새로운 concept를 생성**할 수 있다.   * 이것은 prediction을 바꿀 수 있다. * Feature substitution perturbation은 **현재의 word를 대체하는 새로운 word를 선택**하는 추가적인 단계가 필요하다. * T-EBANO에서는 대체되는 단어와 함께 **이들의 antonym들을 이용**한다. | |

1. **EXPLANATIONS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Explanation은 **Prediction-local explanation, Per class model-global explanation**으로 구분된다.  **< Prediction-local explanation >**   |  |  | | --- | --- | | Textual explanation | Textual explanation은 **모델 분석 결과 가장 관련 있는 feature set을 highlight**한다.   * 많은 feature set은 각각의 interpretable feature 추출 테크닉으로부터 추출될 수 있다. | | Quantitative explanation | 추출된 **feature들의 각각의 set의 영향력을 각 예측에 대해서 분리**하여 보여준다.   * 이때 새롭게 도입된 nPIR index를 이용한다. * 주어진 prediction에 대한 **Input feature의 중요도**를 다음과 같이 계산한다. * **nPIR의 값이 클수록 feature f가 class of interest에 긍정적인 영향** | |

**6. EXPLANATIONS (Cont.)**

|  |
| --- |
| 위 그림은 **textual explanation**의 간단한 예제이다. **BERT model은 텍스트 문서의 sentiment를 탐지할 수 있도록 훈련**되었으며, 각각은 P(positive) 또는 N(negative)이다.   * (a)의 original text에 대해 모델은 negative sentiment를 출력한다. * 사용자는 **(b), (c), (d), (e), (f)에서 highlighted feature들을 검사**하여 어떤 것이 가장 중요한 section인지를 찾아낼 수 있다.     위 표는 위 그림에 대한 **quantitative explanation**을 나타낸다. P는 positive label, N은 negative label이며, (sub.)은 substitution perturbation이 적용되었다는 의미이다.   * 와 는 각각 perturbation 이전과 이후에 모델에 의해 assign된 label을 나타낸다. * Feature set에서 **1개의 feature를 제거하는 것이 모델의 출력을 완전히 다르게 바꿀 수 있으며,** 예측을 (N)에서 (P)로 바꿀 수 있다. * (EXP4), (EXP5)에서 **Substitution perturbation은 적절한 antonym을 찾을 수 있을 때 expressiveness에 큰 영향**을 준다는 것을 알 수 있다.   + (EXP4)에서는 quantitative explanation이 1에 가까운 nPIR 값을 보인다.   + (EXP5)에서는 동사가 부적절한 단어로 교체되어 impact가 없으므로 nPIR은 0이다. |

**6. EXPLANATIONS (Cont.)**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **< Per class model-global explanation >**  T-EBANO는 예측 과정에서 per-class model-global explanation을 제공할 수 있다.   * Input document들에 대한 **local explanation들이 aggregate되고 분석**되며, 모델의 가능한 misleading behavior를 highlight한다. * Global explanation은 각각의 **available class 에 대해 계산**되며, **local explanation set E를 분석**한다.   + 이때 Local explanation 는 **각 문서 와 각각의 interpretable feature f**에 의해 계산된다.   Global influence를 측정하기 위한 다음과 같은 2가지의 index가 있다.   |  |  | | --- | --- | | Global Absolute Influence (GAI) | **Class c에 대해 항상 negative influence**를 준 모든 lemma들에 대해서 0이다.  Class c에 positive한 influence를 주는 각 lemma의 **frequency와 positive influence에 비례하여 증가**한다. | | Global Relative Influence (GRI) | 각각의 class-of-interest에 대해 **가장 영향력이 있고 구별되는 lemma를 highlight**한다. |   각각의 계산 방법은 다음과 같다.   |  |  | | --- | --- | | Global Absolute Influence (GAI) |  | | Global Relative Influence (GRI) | * 어떤 lemma가 **다른 class에 대해 테스트 중인 lemma보다 더 관련성이 클 때** **GRI score = 0**이다. * 테스트 중인 lemma의 영향력이 다른 모든 class에서보다 해당 class c에서 더 클 때 **GRI score > 0**이다. | |

1. **EXPERIMENTAL RESULTS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **7-1. Explanations of the BERT model**     * 각각의 feature extraction 전략이 **최소 1개의 informative local explanation**을 생성한 문서의 비율을 나타낸다. () * Overall은 이들 중 **1개 이상의 전략**이 최소 1개의 informative local explanation을 생성한 문서의 비율을 나타낸다.   **7-2. Explanations of the custom LSTM model**     * 비율은 7-1에서와 같은 의미이다.   **7-3. Textual explanation and Quantitative explanation (Local Explanation)**   |  | | --- | |  | | **Textual explanation report**의 예시이다. 여기서 (a)의 original text는 **LSTM에 의해 원래 Toxic (98%)으로 라벨링**되어 있다. | |  | | 위 그림의 example에 대한 **Quantitiative explanation report**이다. T는 Toxic, C는 Clean 라벨이다. (EXP1: adjective와 noun의 조합, EXP2: MLWE를 이용) | |