**Hate and Toxic Speech Detection in the Context of Covid-19 Pandemic using XAI: Ongoing Applied Research** <https://openreview.net/pdf?id=OLoP-Q2Wu4o>

**0. Explainability and Text Classification**

|  |
| --- |
| 다음과 같은 연구에서 **integrated gradient를 통해 explainability**를 유도한 바가 있다.   * (Sundararajan et al., 2017), (Liu and Avci, 2019)   **SHAP은 더 많이 쓰이는 explainability method**로, feature contribution에 대한 프레임워크를 모델의 출력으로 제공한다.   * 이것은 **cooperative game theory로부터 얻어진 Aultman-Shapely values**에 의해 유도된다. (Lundberg and Lee, 2017) * SHAP의 **Gradient Explainer**는 integrated gradient로 구축되고 expected gradient에 영향을 미친다.   + 이 feature attribution method는 **integrated gradient로부터 적분**을 취하고, 그것을 Shapely value를 계산할 수 있는 expectation으로 변형한다. |

**1. Datasets**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Goal and Basic Info About Dataset:**  **라벨링되지 않은 COVID-19 관련 트윗(robust dataset)에 점수를 매기는** 것이다.   * 이것은 hate and offensive speech의 범위 내에 있는 3가지의 dataset을 합칠 수 있게 한다. 이때 다음의 dataset을 사용한다.   + the collection of racist and sexist tweets presented by Waseem and Hovy (Waseem and Hovy, 2016)   + the Offensive Language Identification Dataset (OLID) (Zampieri et al., 2019)   + Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter (HatEval) (Basile et al., 2019)   **COVID-19 Tweets:**  2개의 서로 다른 source에서 tweet을 수집하였다.   |  |  | | --- | --- | | Texas Advanced Computing Center (TAAC) at the University of Texas at Austin | **“Chinese Virus”가 TACC team이 데이터를 수집하기 위해 사용하는 용어**이기 때문에 중요하다.   * COVID-19에 대한 hate speech의 문맥에서는, 바이러스의 라벨에 있는 국가 또는 민족을 타겟팅하는 용어는 hate speech의 UN guidance를 무시한다. | | Georgia State University’s Panacea Lab | Banda et al., 2020 | |

**2. Experiment Design – Text Classifier**

|  |
| --- |
| Prediction의 성능을 향상시키기 위하여 학습된 feature importance에 영향을 미치는 것이므로, **hate speech classification을 위해 이미 증명된 방법론을 이용**한다. (Gamback and Sikdar, 2017)   * 모든 tweet은 **Glove twitter embeddings**로 변환되며, 이 embedding은 Yoon Kim이 사용한 것과 같은 구조의 **CNN (Convolutional Neural Network) classifier**로 전달된다. (Pennington et al., 2014), (Kim, 2014) * **Random search를 통한 parameter tuning**이 가능하다. * **최종 CNN의 구조**는 다음과 같으며, 파라미터들은 **ranking validation AUC**에 의해 선택되었다.   + **75개의 filter를 갖는 kernel size 3, 4, 6의 CNN**이다.   + 1차원의 max pooling을 적용, dropout rate는 0.4이다.   + Output layer는 l2 regularization 0.029의 softmax layer이다. * 이 모델은 다음과 같이 사용된다.   + baseline model   + XAI를 이용하여 향상된 prediction에 대한 input prediction |

**3. Experiment Design – Calculating Global Average Feature Importance**

|  |
| --- |
| 여기서는 **SHAP의 Gradient Explainer를 baseline model**에 적용한다.   * Gradient Explainer output은 **Glove embedded data와 같은 차원**을 가지므로, 차원을 축소하여 **input text sequence와 일치**시킬 수 있다. * 따라서 **expected gradient를 각 sequence의 용어에 대응되는 axis를 이용하여 합**할 수 있다. |
| **NEXT PAGE** |

**3. Experiment Design – Calculating Global Average Feature Importance**

|  |
| --- |
| **PREVIOUS PAGE** |
| 여기서 **s는 각 sequence에 대한 token**이고, **는 embedding dimension에 대한 모든 expected gradient의 합**이다.   * 이 값은 positive일 수도 있고 negative일 수도 있다. * 이 방법은 **percentage difference를 측정하기 위한 positive input**을 요구하므로, training dataset의 모든 sequence에 대한 모든 s step에서의 이 값들은 **min/max scaling을 통해 0에서 1 사이의 값으로** 바뀐다. * 그 다음에 training corpus로부터 **dictionary**를 만든 후, 각 용어에 대한 **“global average” feature importance**를 저장한다.   + 이 dictionary는 **특정한 prediction이 training data에 나타난 feature importance로부터 얼마나 벗어나 있는지**를 계산하기 위해 사용된다. |

**4. Experiment Design - Enhancing Predictions with Term Difference Multiplier**

|  |
| --- |
| 이제 우리는 각 용어(feature token)의 각 class에 대한 global importance를 알고 있으므로, 각 용어의 **local feature importance와 global feature importance의 percentage difference**를 **각 class에 대해** 계산할 수 있다. 이때 다음 수식을 이용한다.   * 1에서 각 local feature importance와 global feature importance의 percentage difference를 뺀 값을 나타낸다. * 이 값은 sequence에 있는 각 용어에 대해서 **서로 다른 multiplier**를 출력한다. * 이 값들은 **각 sequence에 대해 average**되고, 각 class에 대한 예측된 확률은 그것의 **local Term Difference Multiplier에 의해 multiply**된다.   + 이것은 **local attribution 값이 input sequence에서 각 term의 global mean에 대해 얼마나 떨어져 있는지에 의해 penalty**를 받은, 새로운 **probability score에 대한 예측값**을 출력한다. |

**5. Results and Application to COVID-19 Tweets**

|  |
| --- |
| * Global과 local feature importance score 모두에 대한 **relative neutrality**가 Figure 2에 나타나고, 이것은 **Term Difference Multiplier의 aggregation에 대한 더 높은 overall average**를 유도한다. * 또한, COVID-19 tweet에 이 모델을 적용했을 때, 위와 같이 서로 비슷한 결과를 찾아낼 수 있다.   + **특정한 국가 또는 인종에 대한 정보를 제공하는 tweet은 hateful 또는 toxic**으로 분류된 것을 찾아낼 수 있다. |