삼성카드 데이터 분석 공모전

- KcELECTRA-GCN 기반 고객 피드백 분류 모델 -

TEAM 조진웅

SAMSUNG

전우진
김인조
한재웅

Model Introduction (KcELECTRA-GCN)

➢ 모델 설명 및 선택 사유

- 2021년에 제안된 BertGCN은 GCN의 feature matrix를 BERT로 임베딩된 벡터로 초기화 한 후 GCN layer에서 classification task 수행하며 SOTA를 달성한 바 있음
  - ✓ BertGCN : https://arxiv.org/abs/2105.05727

- 하지만 BERT는 사전학습 효율이 떨어진다는 단점이 있고, 대용량 영어 데이터로 사전학습 된 모델로써 한국어 구어체 데이터에 대해서는 성능의 한계가 있음
  - ✓ 한국어 데이터에 대한 BERT 성능 한계: https://github.com/SKTBrain/KoBERT

- 2020년에 구글에서 발표된 ELECTRA는 BERT의 학습 방식을 개선하며 학습 효율을 개선하였고, 또한 네이버 댓글 및 대댓글로 사전학습하며 한국어 구어체 분류 task에 대해 SOTA를 달성한 KcELECTRA라는 후속 모델이 존재함
  - ✓ KcELECTRA : https://github.com/Beomi/KcELECTRA

- 따라서 본 공모전에서 사용할 모델은 BertGCN에서 Bert 부분을 KcELECTRA로 변경하여 적용하고자 함

- 본 공모전의 task는 한국어 구어체로 구성된 데이터를 적절한 라벨로 분류하는 문제로 KcELECTRA 모델의 한국어 리뷰 데이터 분류 능력과 GCN의 label propagation 능력을 함께 사용하여 정확하게 각 문장을 분류하고자 KcELECTRA-GCN 모델을 고안함

FRAMEWORK

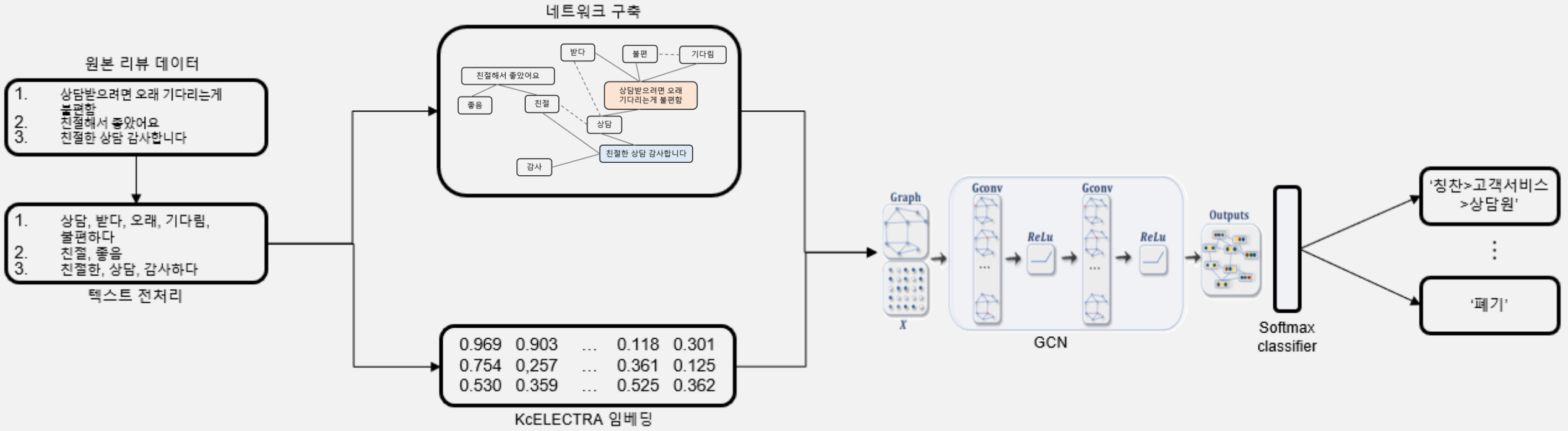

그림 1. 프레임워크

Data Preprocessing

➢ 텍스트 전처리

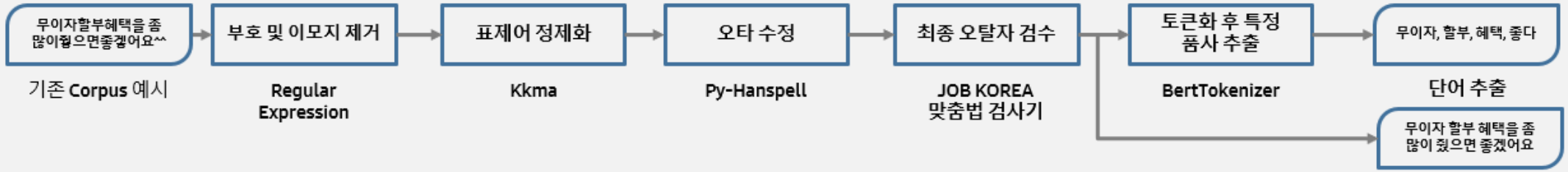

그림 2. 텍스트 전처리

➢ KcELECTRA 임베딩

- 추출된 문장은 KcELECTRA를 이용해 임베딩 진행

➢ 네트워크 구축

- 인접 행렬 (네트워크)
  - ✓ 문장 – 단어를 노드로 하는 two-mode network 구축
  - ✓ '단어-문장' 연결 엣지: TF-IDF weight
  - ✓ '단어-단어'연결 엣지: PPMI weight
- 노드 특성 행렬
  - ✓ 각 노드는 문장 임베딩 벡터를 feature 값으로 가짐
  - ✓ 문장 노드: KcELECTRA representation vector
  - ✓ 단어노드: zero vector

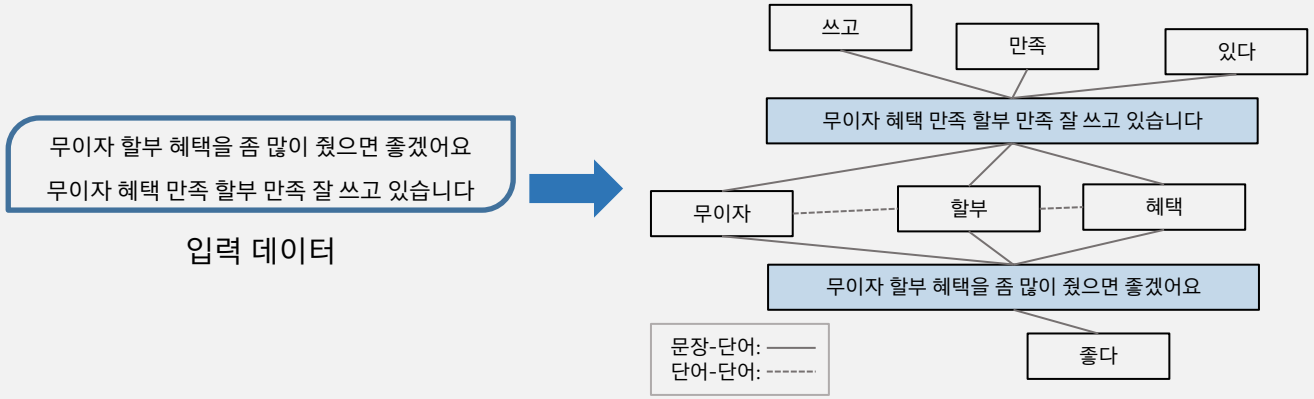

그림 3. 네트워크 구축 예시

Learning Process & Precautions

➢ GCN 학습 과정

- GCN layer는 인접행렬, 각 노드의 KcELECTRA representation 및 라벨을 입력으로 받음
- GCN layer를 거치면서 주변 노드의 정보로 각 노드의 정보가 업데이트됨
- 최종 layer를 거치고 나온 feature vector가 softmax classifier의 입력으로 들어가 라벨 분류 작업 수행

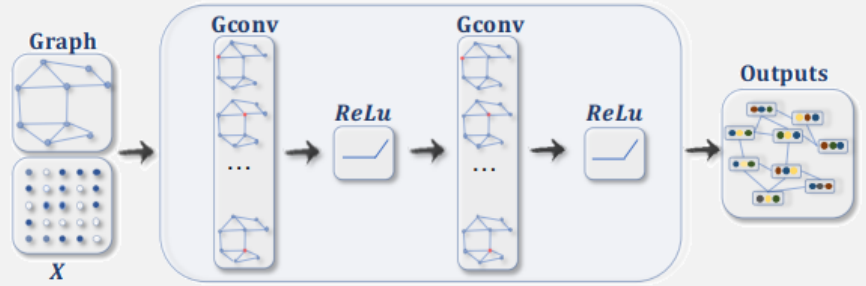

그림 4. GCN 구조

➢ 학습 시 유의사항 및 고려사항

- 복문 처리
  - ✓ Ground-truth를 binary 라벨이 아닌 데이터의 발화의 우선순위를 기준으로 라벨별로 확률을 부여하여(soft label), 복문에서도 유연하게 학습이 가능케 함
- 모델 안정성 향상 방안
  - ✓ 모델의 안정성을 높이기 위해 GCN의 목적 함수와 KcELECTRA의 목적함수의 가중 평균을 최종 목적 함수로 만듦
- 정확도 향상 방안
  - ✓ KcELECTRA-GCN 을 n번 반복하여 출력된 확률 값에 soft voting을 취해 최종 라벨 분류 작업 수행하고자 함
  - ✓ 복문 또는 hand-crafted labeling 과정에서 오분류된 라벨의 경우, 학습과정에서도 명확히 분류되지 않을 수 있으므로 voting을 통해 모델의 신뢰도 및 정확도를 높이고자 함