2023-2 인공지능 응용

이상치 탐지 모델 기반 유튜브 스팸 댓글 탐지

조성수(2018204005), 천흥우(2018311006)

광운대학교 정보융합학부, 산업심리학과

요약

본 논문은 유튜브 뉴스 댓글에서 영상 내용과 무관한 정치 관련 댓글을 탐지하는 모델을 제안한다. 오늘날 유튜브 뉴스를 통해 소식을 손쉽게 접할 수 있고 댓글을 남겨 의견을 공유한다. 하지만, 영상 내용과 무관한 스팸 댓글을 남기는 사람도 있다. 유튜브 자체적으로 스팸 댓글을 관리하지만, 아직 처리되지 않는 댓글이 많다. 기존에도 유튜브 스팸 댓글 탐지하는 연구가 있었다. 스팸 댓글은 정상 댓글 대비 수가 훨씬 적지만, 일반적인 분류모델을 사용하는 경우가 많다. 클래스 불균형 상황인 점을 고려하여, 이상치 탐지 기반 스팸 댓글 탐지 모델을 제안한다. 최근 이슈를 다룬 유튜브 영상 3개에서 수집한 댓글을 활용해, 스팸 댓글 탐지 실험을 수행했다. Isolation Forest, LOF, PCA와 같은 기존의 이상치 탐지 기법과 더불어 딥러닝 기반 이상치 탐지기법을 적용했다.

1. 연구배경

유튜브는 구글에서 제공하는 세계 최대 동영상 플랫폼이다. 과거와는 달리 많은 사람들이 TV가 아닌 유튜브로 뉴스를 접할 수 있다. 유튜브 뉴스 영상에서 많은 누리꾼들이 댓글을 남긴다. 댓글을 통해, 누리꾼 간 소통을 촉진하고 다양한 의견과 생각을 나눌 수 있다. 하지만, 모든 댓글이 긍정적인 역할을 하는 것은 아니다. 혐오, 욕설이나 불건전한 비난은 소셜 네트워크의 분위기를 해칠 수 있다. 그 중에서도 뉴스 영상 내용과 무관하게 특정 정치인이나 자신과 반대되는 정치성향을 가진 사람을 비난하는 댓글이 있다. 이러한 댓글은 건전한 인터넷 사용문화를 저해할 수 있다. 그리고 최혜봉, 김재홍, 이지현, & 이민구 (2020)에 따르면, 정치적으로 편향된 댓글은 객관적인 정보 습득에 부정적인 영향을 줄 수 있다.

그림**1**. 뉴스 영상, 영상 내용과 무관한 스팸 댓글

이러한 이유로 정치 관련 스팸 댓글을 탐지하여 관리할 필요가 있다. 기존에도 스팸 댓글 탐지 모델 연구(논문)이 존재했지만, 한계점이 있다. 스팸 댓글은 정상적인 댓글에 비해 수가 훨씬 적음에도 불구하고 일반적인 분류 모델을 사용한다는 점이다. 클래스 간 불균형 상태에서 일반 분류모델을 사용하는 것은 부적절하다. 불균형 데이터로 분류 모델 학습시, 수가 많은 클래스에 편향되어 학습하게 된다. 결과적으로 정확도는 높지만, 수가 적은 클래스 데이터의 특징은 학습하지 못하게 된다(손민재, 정승원, & 황인준. ,2019). 클래스 간 비중을 균형 있게 하기 위해서, 스팸 댓글을 더 수집하면 되지만, 시간과 비용이 더 많이 들게 된다. 본 논문에서는 이상치 탐지 모델 기반 스팸 댓글을 제작함으로써, 기존 방식의 한계점을 개선하고자 한다.

2. 문제 설명 및 데이터 소개

이상 정의: 스팸 댓글은 뉴스 영상 내용과 무관한 정치 관련 댓글로 정의한다.
문제 정의: 임베딩 벡터로 스팸 댓글을 어떻게 탐지할 수 있는가?

유튜브 API를 사용하여, 정치와 무관한 3개의 영상에서 18,178개의 댓글 수집했다. html 태그와 특수문자를 제거하고 띄어쓰기 오류를 교정하는 등의 전처리를 했다. 한글이 없는 댓글을 필터링하여 총 댓글 17,547개를 실험에 사용했다. 정치 관련 단어 사전을 만들어서, 댓글에 사전에 있는 단어가 포함된 경우

스팸댓글(이상치, 1)로 라벨링 했다. 반대로, 댓글에 사전에 포함된 단어가 없는 경우 정성 댓글(정상치, 0)로 라벨링 했다. 임베딩은 단어 수준 임베딩(tf-idf)과 문장 수준 임베딩(SentenceTransformer)을 하였다. 문장 임베딩의 경우, 사전 학습된 언어 모델 SBERT, KCBERT, KCELECTRA를 활용했다. 768차원으로 임베딩 차원이 많기 때문에, 5차원으로 차원축소(PCA)했다. TF-IDF도 5차원으로 차원축소를 하였다.

3. EDA

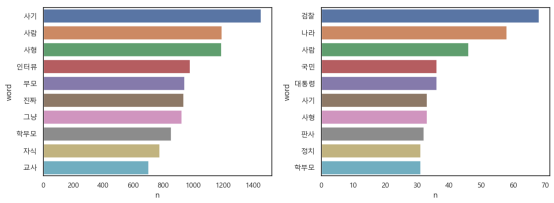

그림**2.** 정상 댓글과 스팸 댓글의 빈도순 상위 10개 단어 막대 그래프.

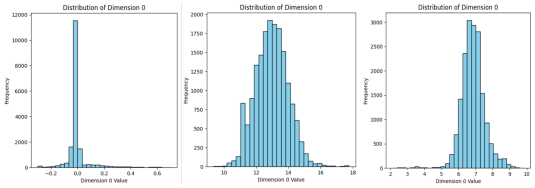

그림**3.** (왼쪽부터) TF-IDF, KCBERT, KCELECTRA의 첫번째 임베딩 차원의 히스토그램

정상댓글(0)은 17,129개이고, 스팸 댓글(1)은 418개이다. 차원 축소에 의해 5개의 독립변인이 존재한다. TF-IDF의 경우, 각 임베딩 차원의 최소값이 -0.4~-0.2이고 최대값은 0.6~0.8이다. 각 차원마다 최대, 최소값의 편차가 존재한다. 그리고 각 차원의 분포가 왼쪽으로 쏠리는 경향이 있었다. SentenceTransformer(KCBERT)의 경우, 각 차원의 최소값이 -8~-10이고 최대값은 6~18이다. SentenceTransformer(KCELECTRA)의 경우, 각 차원의 최소값이 -4~2이고 최대값은 4~10이다. SentenceTransformer을 통해 얻은 임베딩 차원은 TF-IDF보다 정규분포에 가까운 형태를 띈다.

4. 모델 학습

SMOTETomek, TomekLinks를 이용하여 데이터 전체의 정상 댓글과 스팸 댓글 간 불균형 문제를 해소해주었다. 그리고 학습 데이터에서, 각 임베딩 차원마다 스팸 댓글과 정상댓글에 가중치를 부여했다. 가중치는 Scikit learn 라이브러리에서 제공하는 class_weight 함수를 이용했다. Isolation Forest, LOF, PCA, Auto Encoder 기반 이상치 탐지 모델을 사용했다. 그리고 추가로 뉴럴 네트워크 기반 이상치 탐지 기법 CNN, XGBOD, DeVNet 3가지를 제시한다. DeVNet의 경우, 추가로 ablation study를 진행했다.

4.1.1. CNN

이상치 탐지 모델로 사용되는 CNN은 기존 연구에서 제시되었으며 다음과 같은 특징을 갖는다(Kwon et al, 2018). CNN은 원래 이미지 분석을 위해 설계되었지만, 텍스트 데이터에서도 공간적 특징 추출 능력을 효과적으로 활용할 수 있다. 텍스트에서의 "공간"은 주로 단어 및 문맥의 배열을 나타내며, CNN은 이를 통해 각 단어의 주변 문맥을 학습하고 특징을 추출할 수 있다. 문맥 파악과 문장 내 중요한 특징 자동 추출이 가능하다.

4.1.2. XGBOD

XGBOD (Extreme Gradient Boosting for Outlier Detection)는 이상 탐지에 사용되는 강력한 알고리즘 중 하나이다. 그 중에서, XGBoost라는 트리 기반의 그래디언트 부스팅 알고리즘이 있다(Yan, Z., et al, ,2023). 각 데이터 포인트에 대해 이상치 스코어를 계산한다. 그리고 각 트리에서 어떤 특성이 이상치 탐지에 큰 영향을 주었는지 파악할 수 있다.

4.1.3. DevNet (Deviation Network)

DevNet은 기존 연구(Pang, G., et al, 2019)에서 제시되었고, 다음과 같은 특징을 갖는다. 비선형 패턴을 탐지하는데 성능이 뛰어나다. 그리고 계층적인 특징을 학습할 수 있다. L2 정규화를 사용하여 모델의 일반화 성능을 향상시키는 동시에, 복잡한 패턴 적합 성능도 높다.

5. 실험 결과 및 결론

Isolation Forest, LOF, PCA, Auto Encoder는 대부분의 관측치를 정상 댓글 또는 스팸 댓글로 판별하는 경향성이 있었다. 뉴럴 네트워크 기반 이상치 탐지 모델을 사용한 경우, 이상치와 정상 댓글을 비교적 정확하게 탐지했다. 그 중에서도 DCNN, XGBOD가 월등히 높은 정확도를 보였다.

5.1 Ablation Study(DevNet): 정규화 강도 조절

TF-IDF를 이용한 DevNet 모델의 L2 정규화 수준을 0.01에서 0.005로 줄여보았다. 그리고 모델 성능의 변화를 살펴보았다. 정확도가 57.47%로 근소하게 증가했고, Precision이 73.77%에서 64.76%로 감소했다. 그리고 Recall이 92.99%에서 82.66%로 큰 폭으로 감소했다. 정규화 수준에 따라 Recall과 Precision이 민감하게 변함을 알 수 있다.

5.2 Ablation Study(DevNet): 32 Dense Layer 제거

Dense layer 32개를 제거하여, 모델 성능의 변화를 확인해보았다. Precision이 73.77%에서 65.41%로, Recall이 92.99%에서 83.99%로 감소했다. Dense layer 제거 여부에 따라서 모델 성능의 큰 차이를 보였다.

Model	Embedding	Dimension	Accuary	Recall	Precison	F1 Score
Isolation Forest	KCBERT	5	41.7%	3.8%	16.8%	0.062
LOF	KCELECTRA	5	47.5%	79.9%	49%	0.607
PCA	KCELECTRA	5	47%	13%	42.7%	0.199
Auto Encoder	KCELECTRA	5	48.5%	15.7%	48%	0.236
AE + IForest	KCELECTRA	5	46%	11%	39.9%	0.172
AE + LOF	KCBERT	5	50.8%	14.6%	55.4%	0.231
SCNN	KCELECTRA	5	68.97%	75.95%	82.48%	0.72
DCNN	KCELECTRA	5	**90.72%**	**95.4%**	**95.88%**	**0.91**
DEVNET	TF-IDF	5	56.55%	73.77%	92.99%	0.58
XGBOD	KCELECTRA	5	**97.2%**	**96.7%**	**96.67%**	**0.97**

표**1.** 스팸 댓글 탐지 모델 성능 비교

6. 참고문헌

Kwon, D., Natarajan, K., Suh, S. C., Kim, H., & Kim, J. (2018, July). An empirical study on network anomaly detection using convolutional neural networks. In *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)* (pp. 1595-1598). IEEE.

Pang, G., Shen, C., & van den Hengel, A. (2019, July). Deep anomaly detection with deviation networks. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 353-362).Yan, Z., Sun, J., Yi, Y., Yang, C., & Sun, J. (2023). Data-Driven Anomaly Detection Framework for Complex Degradation Monitoring of Aero-Engine. *International Journal of Turbomachinery, Propulsion and Power*, *8*(1), 3.

손민재, 정승원, & 황인준. (2019). 불균형 데이터 분류를 위한 딥러닝 기반 오버샘플링 기법. 정보처리학회논문지. 소프트웨어 및 데이터 공학, *8*(7), 311-316.

최혜봉, 김재홍, 이지현, & 이민구. (2020). 정보 중립성 확보를 위한 인터넷 뉴스 댓글의 정치성향 분석. *The Journal of the Convergence on Culture Technology (JCCT)*, *6*(4), 575-582.

댓글 수집에 사용된 영상

[단독]전청조 첫 방송 인터뷰 “죗값 받겠다” | 뉴스

[자막뉴스] "넌 자식이 없어서 그러냐?" 폭언하던 학부모, 결국... / YTN

'부산 서면 돌려차기 사건' 드디어 밝혀진 진실 [shorts]