**Performance Evaluation and Assessment Methods for Keyword Extraction Models**

**1. Introduction**

키워드 추출은 텍스트 데이터에서 중요하고 관련성 높은 단어나 구를 자동으로 식별하는 과정을 의미합니다. 현대 사회에서 텍스트 데이터의 양이 폭발적으로 증가함에 따라, 이를 효율적으로 처리하고 분석하는 기술의 중요성도 함께 증가하고 있습니다. 키워드 추출은 이러한 대량의 텍스트 데이터를 요약하고 주요 정보를 빠르게 파악하는 데 핵심적인 역할을 합니다. 이는 문서의 주요 내용과 주제를 이해하는 데 도움을 주며, 텍스트 요약, 검색 엔진 최적화(SEO), 정보 검색, 문서 분류 등 다양한 분야에 활용될 수 있습니다. 특히, 검색 엔진 최적화와 같은 분야에서는 적절한 키워드를 추출하여 웹페이지의 가시성을 높이는 데 중요한 역할을 합니다. 또한, 정보 검색 시스템에서는 키워드 추출을 통해 사용자가 원하는 정보를 신속하고 정확하게 찾을 수 있도록 합니다.

본 연구의 목적은 다양한 키워드 추출 모델의 성능을 평가하고, 각 모델의 강점과 한계를 분석하는 것입니다. 키워드 추출 모델은 크게 통계 기반 모델과 딥러닝 기반 모델로 나눌 수 있습니다. 통계 기반 모델은 단어의 빈도와 같은 통계적 특성을 이용하여 키워드를 추출하는 반면, 딥러닝 기반 모델은 문맥과 단어 간의 의미적 관계를 고려하여 키워드를 추출합니다. 이번 연구에서는 TF-IDF, TextRank와 같은 통계 기반 모델과 KeyBERT, GPT-3.5, GPT-4와 같은 딥러닝 기반 모델의 성능을 비교하고자 합니다. 이를 통해 각 접근법의 효율성을 평가하고, 더 나아가 키워드 추출 모델 평가 방식을 제안하여 그 성능을 검토하고자 합니다.

본 연구의 기여점은 다음과 같습니다. 1. 키워드 추출 모델들의 특징 및 성능 비교. 2. 키워드 추출 모델을 평가할 모델 파이프라인을 제안.

**2. Literature Review**

키워드 추출은 문서에서 주요 단어나 구를 식별하여 문서의 내용을 요약하거나 특정 주제를 빠르게 파악할 수 있게 하는 중요한 자연어 처리(NLP) 기술입니다. 키워드 추출 방법은 크게 두 가지 접근법으로 나눌 수 있습니다: 통계 기반 접근법과 딥러닝 기반 접근법입니다.

**2.1. 통계 기반 방법**

**TF-IDF (Term Frequency-Inverse Document Frequency) :** TF-IDF는 문서 내에서 특정 단어의 중요성을 평가하는 통계적 방법입니다. TF(Term Frequency)는 특정 문서에서 단어가 등장하는 빈도를 나타내고, IDF(Inverse Document Frequency)는 단어가 전체 문서 집합에서 얼마나 자주 등장하는지를 나타냅니다. TF-IDF는 이 두 값을 곱하여 계산되며, 특정 문서에 많이 등장하지만 전체 문서에서는 드물게 나타나는 단어를 중요한 키워드로 식별합니다. TF-IDF는 구현이 간단하고 효율적이지만, 단어의 순서나 문맥을 고려하지 않기 때문에 의미론적 유사성을 반영하지 못하는 단점이 있습니다. **TextRank** : TextRank는 그래프 기반의 키워드 추출 방법으로, 단어 또는 구 사이의 관계를 그래프로 모델링합니다. 각 노드는 키워드 후보를 나타내고, 간선은 이들 사이의 의미적 또는 문법적 관계를 나타냅니다. PageRank 알고리즘을 사용하여 각 노드의 중요도를 계산한 후, 상위 노드를 키워드로 선택합니다. TextRank는 단어 간의 관계를 고려하여 문맥을 반영할 수 있다는 장점이 있지만, 긴 문서의 경우 계산 비용이 높아질 수 있습니다.

**2.2. 딥러닝 기반 방법**

**KeyBERT** : KeyBERT는 사전 학습된 BERT(Bidirectional Encoder Representations from Transformers) 모델을 사용하여 키워드를 추출합니다. N-gram에 대한 임베딩을 추출하고, 주어진 데이터와의 코사인 유사도를 계산하여 키워드를 선택합니다. KeyBERT는 문맥을 고려할 수 있고, 하이퍼파라미터로 다변성을 조절하여 유사한 키워드의 반복을 방지할 수 있습니다. 그러나 사전 학습된 BERT 모델의 성능에 크게 의존하며, 계산 비용이 높다는 단점이 있습니다. **GPT-3.5와 GPT-4** : GPT(Generative Pre-trained Transformer)는 대규모 언어 모델로, 다양한 NLP 작업을 처리할 수 있습니다. GPT-3.5와 GPT-4는 transformer 아키텍처를 기반으로 하며, 광범위한 텍스트 데이터를 사전 학습하여 높은 수준의 언어 이해 및 생성 능력을 갖추고 있습니다. 이 모델들은 키워드 추출에서도 강력한 성능을 보이며, 특히 문맥을 깊이 있게 이해하고 중요한 키워드를 식별하는 데 뛰어납니다. 하지만 많은 연산 자원을 필요로 하며, 비용이 많이 듭니다.

통계 기반 방법과 딥러닝 기반 방법은 각기 다른 장단점을 가지고 있습니다. 통계 기반 방법은 구현이 간단하고 빠르게 실행되지만, 문맥을 고려하지 못하는 한계가 있습니다. 반면, 딥러닝 기반 방법은 높은 정확도와 문맥 이해 능력을 제공하지만, 계산 비용이 높고 복잡한 구현이 필요합니다.

**3. Methods**

**3.1. Dataset**

본 연구에서는 한국어 기사를 대상으로 키워드 추출 실험을 진행하였습니다. 데이터를 확보하기 위해 국내 최대 포털 사이트인 네이버에서 정치, 엔터테인먼트 등 다양한 주제의 기사를 크롤링하였습니다. 각 기사에서 핵심 문단을 발췌하여 총 30개의 데이터셋을 구성하였습니다. 이러한 접근법을 통해 다양한 주제와 문체를 포함하는 데이터를 수집하였습니다. 데이터 전처리는 Unigram 토크나이저를 사용하여 각 문장을 띄어쓰기를 기준으로 토큰화하는 방식으로 진행하였습니다. 명사를 따로 분리하거나 조사와 같은 불용어를 제거하지 않고 원문 그대로를 유지함으로써 실험의 일관성을 확보하였습니다.

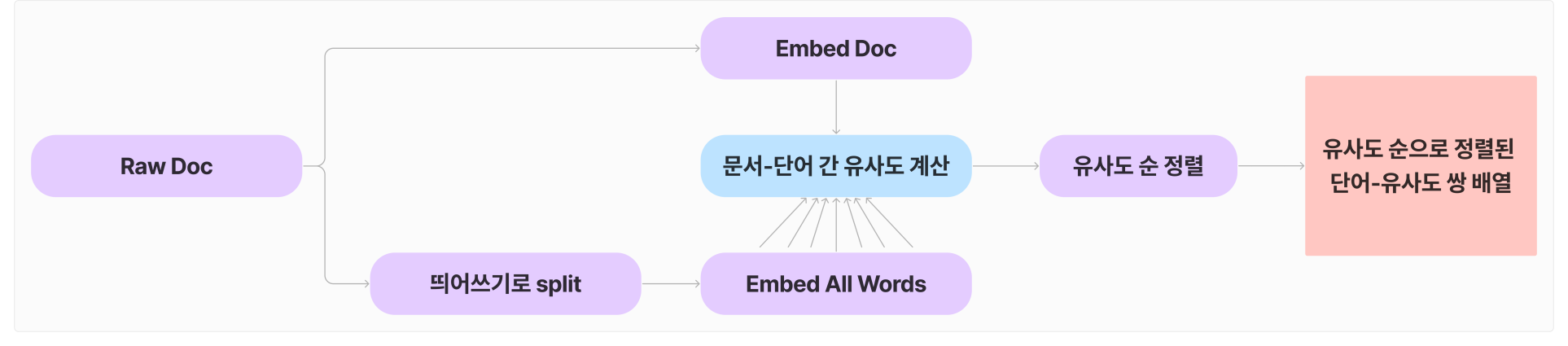
**3.2. Model**

**TF-IDF :** TF-IDF 모델은 각 단어의 등장 빈도(TF)와 역문서 빈도(IDF)를 계산하여 중요 단어를 추출합니다. **TextRank :** TextRank 모델은 단어 간의 관계를 그래프로 모델링하여 키워드를 추출합니다. 각 단어를 노드로 하고, 단어 간의 유사도를 간선으로 하는 그래프를 구성합니다**. KeyBERT :** KeyBERT 모델은 사전 학습된 BERT 임베딩을 사용하여 문맥을 고려한 키워드 추출을 수행합니다**. GPT-3.5와 GPT-4 :** GPT-3.5와 GPT-4 모델은 대규모 사전 학습된 언어 모델로, 문서의 문맥을 깊이 있게 이해하여 키워드를 추출합니다.

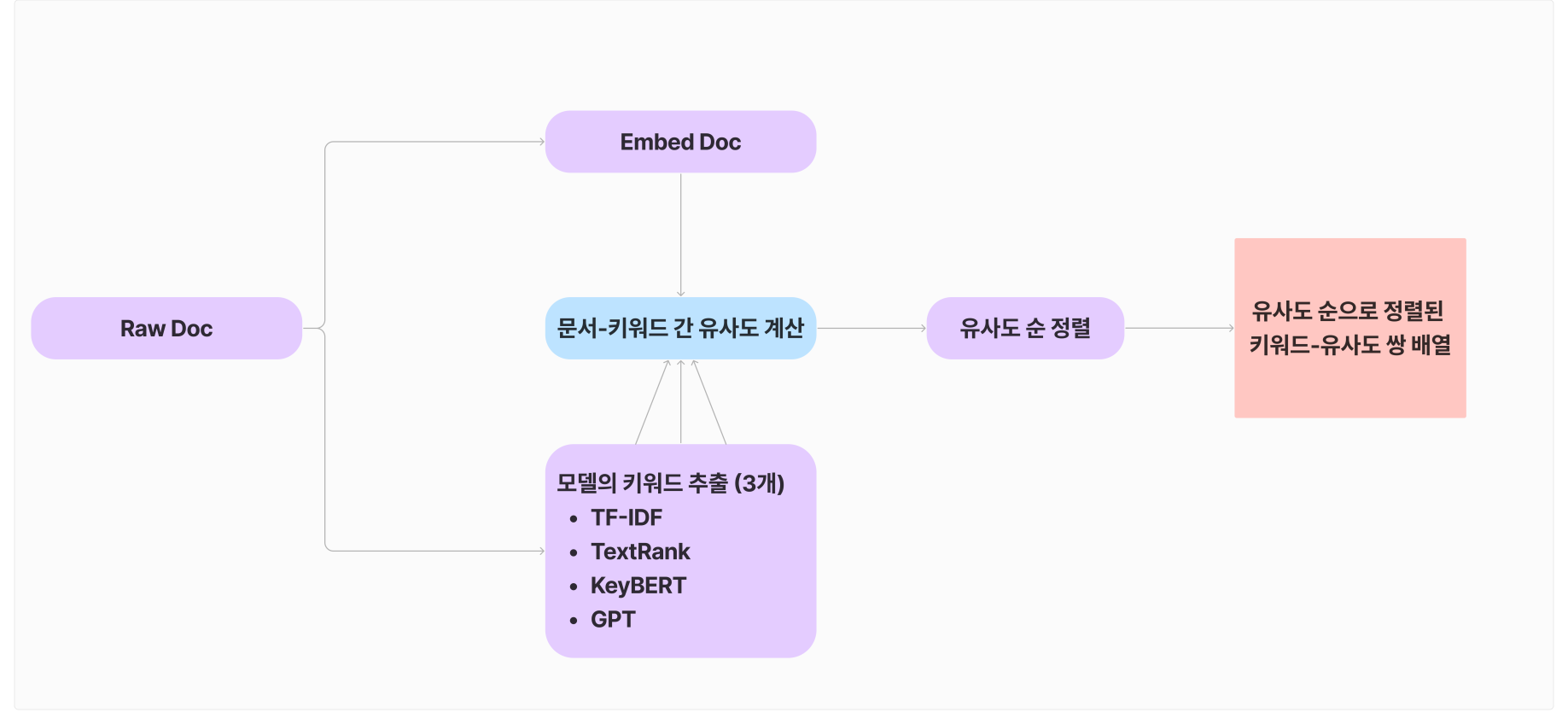
**3.3. Experiment Setup**

각 모델은 입력된 문단으로부터 3개의 핵심 키워드를 추출하도록 설정하였습니다. 실험의 일관성을 위해 모든 모델에 동일한 데이터셋을 사용하였으며, 모델의 성능을 객관적으로 평가하기 위해 코사인 유사도를 기준으로 평가하였습니다.

**3.4. Evaluation Model**

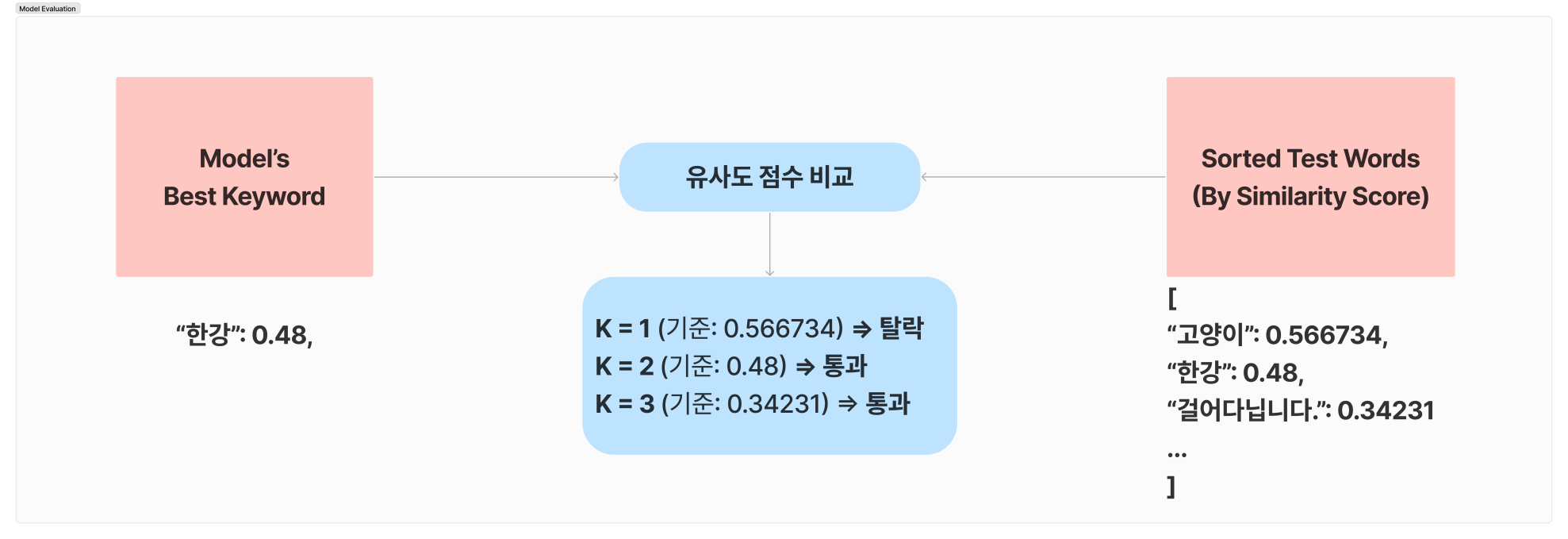


**Fig 1-(a).** 평가용 단어 – 유사도 셋 구축



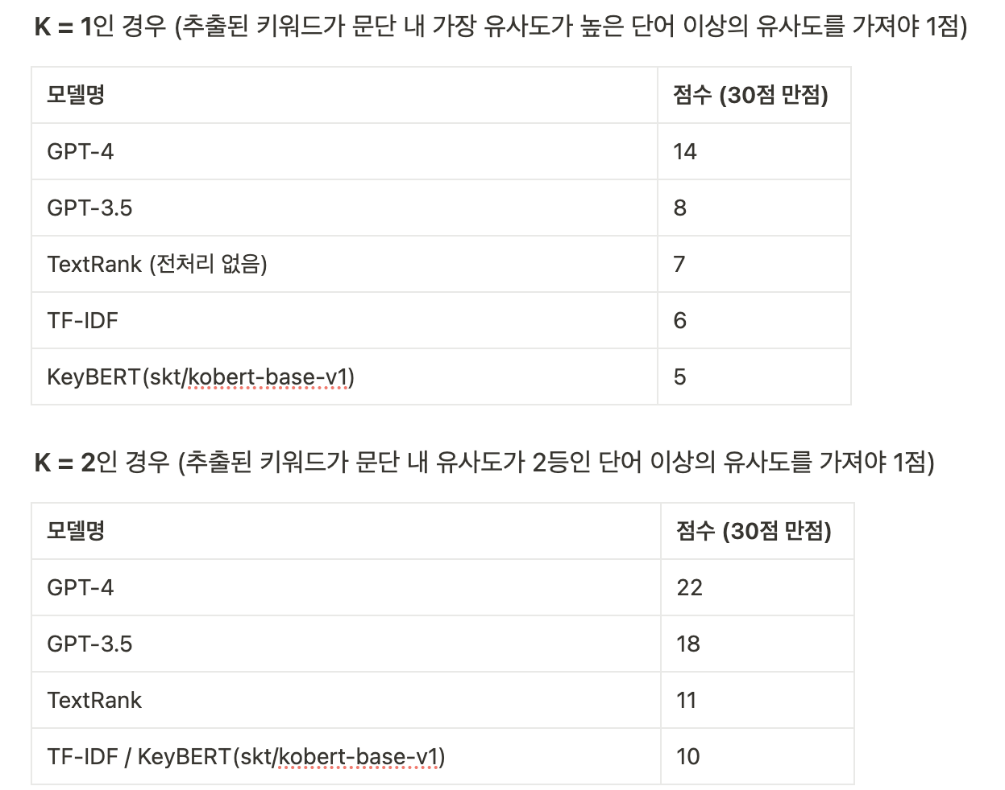
**Fig 1-(b).** 모델 키워드 추출

**Fig1-(c).** 모델 평가

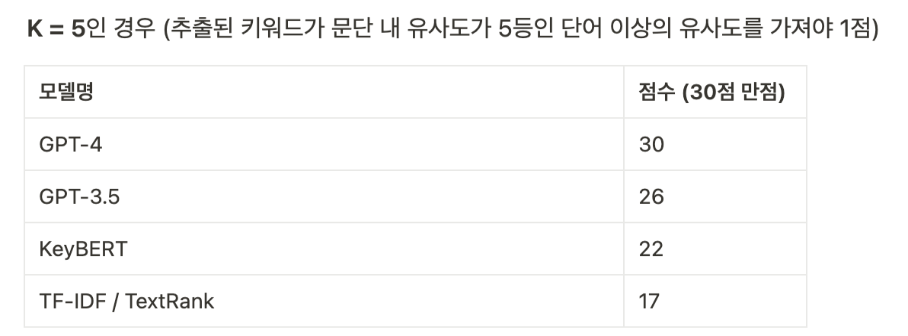


모델의 성능 평가는 각 모델이 추출한 키워드와 문단 간의 코사인 유사도를 계산하여 이루어졌습니다. 코사인 유사도는 두 벡터 간의 유사성을 측정하는 척도로, -1에서 1 사이의 값을 가지며, 값이 클수록 유사성이 높음을 나타냅니다. 모델이 추출한 3개의 키워드 중 해당 키워드와 문단 사이의 코사인 유사도가 가장 높은 단어 하나를 선정합니다. 이와 함께 문단 내에서 해당 문단을 이루는 각 단어와 전체 문단 사이의 코사인 유사도를 모두 구합니다. 이를 내림차순으로 정렬하여 상위 k개의 점수와 모델이 선정한 키워드의 점수를 비교합니다. 만약 모델이 추출한 키워드의 유사도 점수가 k번째 기준 단어의 점수보다 더 높다면 1점을 부여합니다. 이런 방식으로 각 문단마다 한 번씩, 총 30점 만점으로 테스트를 수행합니다. 이 점수는 해당 모델이 키워드를 얼마나 잘 추출했는지에 대한 척도로 해석이 가능합니다. 해당 방법은 모델 자체의 성능뿐만 아니라, 모델 간 성능을 쉽게 비교할 수 있다는 장점이 있습니다.

**4. Experiment**



실험 결과 GPT4 모델이 가장 높은 성능을 보였습니다. 이는 GPT4가 문단의 가장 중요한 내용을 가장 잘 포착한다는 것을 의미합니다. K가 2,3,4,5로 늘어남에 따라 모든 모델의 점수가 전반적으로 증가하였는데, 이는 k값의 증가로 인한 기준 완화 때문이라고 해석할 수 있습니다. 특히 k가 5일 때 GPT4가 만점을 획득하였으며, GPT4가 모든 시나리오에서 주요 키워드를 잘 추출하고 있음을 나타냅니다. 실험 결과를 종합적으로 해석해보면 GPT4는 일관되게 높은 성능을 보이며 복잡하고 다양한 문맥에서도 키워드를 잘 추출하는 능력을 가진다고 평가할 수 있습니다. GPT3.5는 GPT4에 비해서는 다소 떨어지지만, 충분히 성능이 우수한 모델이라고 평가할 수 있습니다. KeyBERT의 경우 딥러닝 모델이지만 GPT보다 낮은 성능을 보였습니다. 이는 딥러닝 모델의 성능에 있어서 임베딩 모델의 역할이 중요함을 보여줍니다. 통계 기반의 TextRank, TF-IDF 모델은 비교적 성능이 떨어지는 것으로 보이는데, 이는 통계 기반 모델이 문맥을 파악하지 못하는 점 때문인 것으로 판단됩니다.



**5. Limitations**

**5.1. 데이터셋의 크기**

본 연구에서는 시간과 자원의 제한으로 인해 30개의 기사 데이터셋을 사용하여 실험을 진행하였습니다. 비록 이 데이터셋이 다양한 주제를 포함하고 있지만, 데이터의 크기가 상대적으로 작아 모델 간의 성능 차이를 명확히 파악하는 데 한계가 있었습니다. 더 많은 데이터셋을 사용하여 실험을 수행하였다면, 결과의 신뢰성과 일반화 가능성을 높일 수 있었을 것입니다. 특히, K=1일 때 GPT-3.5, TextRank, TF-IDF의 성능 차이가 근소하게 나타났는데, 더 큰 데이터셋을 사용하면 이러한 모델 간의 성능 차이를 더욱 명확하게 확인할 수 있었을 것입니다.

**5.2. 문단 길이의 다양성 부족**

실험에 사용된 입력 문단의 평균 길이는 약 218개의 글자와 49개의 단어로 구성된 비교적 짧은 수준의 문단들입니다. 실험 결과의 일관성을 유지하기 위해 균일한 길이의 문단을 사용했으나, 문단 길이에 따른 모델 성능 변화를 평가하지 못했습니다. 다양한 길이의 문단을 포함한 데이터셋을 사용하여, 짧은 문단, 중간 길이의 문단, 긴 문단에서의 모델 성능을 비교해 보는 것이 의미가 있을 것입니다.

**5.3. 전처리 과정의 제한**

이번 연구에서는 Unigram 토크나이저를 사용하여 각 문장을 띄어쓰기를 기준으로 토큰화하였습니다. 명사를 따로 분리하거나 조사와 같은 불용어를 제거하지 않았기 때문에, 이러한 전처리 과정의 부재가 키워드 추출 성능에 영향을 미쳤을 수 있습니다. 더욱 정교한 형태소 분석과 불용어 제거를 통해 전처리 과정을 개선하면, 모델의 성능을 향상시킬 수 있을 것입니다.

**5.4. 검증 데이터셋의 부재**

한글 키워드 추출 모델의 성능을 평가할 수 있는 공인된 검증 데이터셋을 확보하지 못했습니다. 자체적으로 크롤링한 뉴스 기사 데이터를 사용하여 실험을 진행하였기 때문에, 평가 결과가 주관적일 수 있습니다. 공인된 검증 데이터셋이 있었다면, 모델 성능을 보다 객관적으로 평가할 수 있었을 것입니다. 따라서, 향후 연구에서는 검증 데이터셋 확보를 통해 평가의 신뢰성을 높이는 것이 필요합니다.

이러한 한계에도 불구하고, 본 연구는 다양한 키워드 추출 모델의 성능을 비교하고, 각 모델의 강점과 약점을 파악하고 모델 성능 평가 파이프라인을 구축하는데 의미 있는 기여를 하였습니다.

**6. Conclusion**

본 연구는 다양한 키워드 추출 모델의 성능을 평가하고, 각 접근법의 장단점을 분석하는 데 중점을 두었습니다. 통계 기반 모델인 TF-IDF와 TextRank, 딥러닝 기반 모델인 KeyBERT, GPT-3.5, GPT-4를 비교하여, 한국어 기사 데이터셋을 대상으로 키워드 추출 성능을 평가하였습니다. 평가 결과는 코사인 유사도를 기반으로 하여, 각 모델이 추출한 키워드가 원 문단과 얼마나 유사한지를 측정하였습니다. 본 연구는 다양한 키워드 추출 모델을 비교함으로써, 각 모델의 강점과 한계를 명확히 파악하는 데 기여하였습니다. 또, 키워드 추출 모델의 평가방식을 고안하여 성능 평가 파이프라인을 제안하는데 기여하였습니다.

**7. Future Works**

**더 큰 데이터셋 사용 :** 향후 연구에서는 더 많은 데이터를 사용하여 모델 간의 성능 차이를 더욱 명확하게 파악할 것입니다. 다양한 길이의 문단을 포함한 데이터셋을 통해 문단 길이에 따른 모델 성능 변화를 평가할 필요가 있습니다.

**정교한 전처리 과정 :** 형태소 분석과 불용어 제거 등 정교한 전처리 과정을 도입하여 모델 성능을 향상시킬 수 있는 방안을 모색할 필요가 있습니다.

**검증 데이터셋 확보 :** 공인된 한글 키워드 추출 데이터셋을 확보하여 보다 객관적이고 신뢰성 있는 평가를 수행할 필요가 있습니다.

**8. References**

Salton, G; McGill, M. J. (1986). *Introduction to modern information retrieval*. McGraw-Hill. ISBN 978-0-07-054484-0.

Mihalcea, R., & Tarau, P. (2004). TextRank: Bringing order into text. In D. Lin & D. Wu (Eds.), Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (pp. 404–411). Association for Computational Linguistics. <https://aclanthology.org/W04-3252>

Brown, T. B., et al. (2020). *Language models are few-shot learners*. arXiv. <https://arxiv.org/abs/2005.14165>

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding.* arXiv. <https://arxiv.org/abs/1810.04805>