## 텍스트 마이닝 과제

### 배경

부동산 시장에 대한 관심이 여전히 높다는 점은 여러 경제 지표와 사회적 흐름을 반영하고 있습니다. 최근 금리가 상승하지만, 많은 투자자들은 부동산을 안정적인 자산으로 보고있습니다. 대환대출과 주택담보대출 갈아타기를 통해 금융 부담을 줄이려는 움직임이 활발해지면서 이러한 경향이 더욱 뚜렷해지고 있습니다.

한국은행의 기준금리 인상에 따라 주택담보대출 이자율이 상승하고 있으며, 이는 주택 구매자나 투자자에게 부담이 될 수 있습니다. 그러나 대환 대출을 통해 기존 대출의 이자 부담을 줄이려는 수요가 증가하고 있습니다. 이와 함께 부동산 매물의 공급과 수요 변화도 중요한 요소입니다. 매물 수가 줄어들 경우 가격 상승 가능성이 높아지고 반대로 공급이 많아지면 가격 하락 압력이 커질 수 있습니다.

또한, 정부의 부동산 관련 정책이나 규제는 시장에 큰 영향을 미치며, 세금 정책이나 대출 규제의 변화가 투자자들의 심리에 미칠 수 있습니다. 소비자와 투자자의 심리적 요인도 간과할 수 없습니다.

이러한 요소들을 종합적으로 분석하여 부동산 기사를 스크랩하고 분석하면, 이진 분류 모델을 통해 시장의 현황을 신속하게 판단할 수 있습니다. 기사 내용에서 긍정적인 신호와 부정적인 신호를 분류하여 시장 방향성을 예측하는데 도움을 줄 수 있고, 딥러닝을 활용하여 시장 변화를 예측하는 모델을 구축하여 방향성을 예측할 수 있습니다.

### 데이터 설명

제목 : 기사 제목,

본문 : 기사 내용,

TEXT : 제목과 본문 병합한 텍스트 데이터,

감성점수 : TEXT 엔티티 행의 단어별 긍정,부정 도합 점수,

lama 3.1 감성 분석 : Ollama를 사용한 감성분석 결과

### 방법론

1. **데이터 수집**: Selenium과 BeautifulSoup를 활용하여 부동산 관련 뉴스를 웹에서 크롤링합니다. 동적 웹 페이지를 처리하기 위해 undetected\_chromedriver를 사용합니다.
2. **데이터 전처리**: 수집한 데이터를 정리하여 텍스트 결합, 명사 추출 등의 전처리 작업을 수행합니다. KoNLPy의 Okt 모듈을 사용하여 텍스트에서 명사를 추출합니다. 모델링을 위해 TEXT 엔티티를 TF-IDF 벡터화, y값을 수치화합니다.
3. **분석 방법**: 명사 빈도 분석, 감성 분석 및 고급 머신러닝 모델(T5, Llama3.1)을 통해 텍스트 데이터를 분석합니다. ->변경 AutoML로 적합한 머신러닝 파악 ->LGBM통해 텍스트 데이터 훈련 후 감성 사전을 구축하여 뉴스 기사의 긍정, 부정, 중립을 분류합니다.
4. **결과 시각화**: 빈도 분석 결과를 시각화하여 주요 단어를 한눈에 확인할 수 있도록 합니다. Matplotlib와 WordCloud를 활용합니다.

### 분석 과정

1. **데이터 수집**:
   * 웹 크롤링을 통해 부동산 뉴스 제목과 본문을 수집합니다.
   * undetected\_chromedriver를 사용하여 동적 콘텐츠를 로드합니다.
2. **데이터 전처리**:
   * 수집한 뉴스 제목과 본문을 결합하여 하나의 텍스트 필드를 생성합니다.
   * KoNLPy의 Okt 모듈로 텍스트에서 명사를 추출하고, 2글자 이상의 명사만 필터링합니다.
3. **빈도 분석**:
   * 추출된 명사의 빈도를 계산하고, 상위 20개 단어를 선정합니다.
   * Matplotlib을 사용하여 빈도수를 시각화하고, WordCloud를 생성하여 중요 단어를 강조합니다.
4. **감성 분석**:
   * KNU 감성 사전을 기반으로 긍정, 부정, 중립의 감성을 판단하는 사전을 구축합니다.
   * 각 뉴스 텍스트에 대해 감성 점수를 계산하고 데이터프레임에 추가합니다.
5. **고급 감성 분석**:
   * OpenAI의 Llama3.1 모델을 사용하여 각 뉴스의 감성을 분석합니다.
   * 분석 결과를 데이터프레임에 추가하고 최종 결과를 저장합니다.
6. **모델 훈련 및 예측**:
   * Hugging Face의 T5 모델을 사용하여 텍스트 분류 모델을 훈련합니다.
   * 훈련된 모델을 통해 새로운 텍스트의 감성을 예측합니다.

### 결론

텍스트, 스크린샷, 도표, 번호이(가) 표시된 사진

자동 생성된 설명텍스트, 폰트, 스크린샷, 일렉트릭 블루이(가) 표시된 사진

자동 생성된 설명

**1. 빈도 분석 결과**

첫 번째 그래프에서 나타난 바와 같이, **아파트와 주택**이라는 키워드는 가장 높은 빈도를 기록하고 있습니다. 이는 현재 부동산 시장에서 아파트와 주택이 주요 관심사임을 의미합니다. 특히, **서울**과 **분양** 관련 키워드도 높은 빈도로 나타나, 서울 지역 내 아파트 및 주택 분양에 대한 수요가 여전히 강하다는 것을 보여줍니다.

* **아파트**: 주거 형태 중 가장 많은 관심을 받고 있으며, 이는 안전한 투자처로 인식되고 있기 때문일 수 있습니다.
* **주택**: 아파트 외에도 다양한 주택 유형에 대한 관심이 높아지고 있으며, 이는 대출 및 주택 구입 관련 정책 변화에 따라 더욱 부각되고 있습니다.
* **서울**: 서울 지역의 부동산 시장이 특히 주목받고 있으며, 이는 수도권 집중 현상과 관련이 있을 수 있습니다.
* **분양**: 신규 아파트 분양에 대한 관심이 높아, 이는 시장의 공급 측면에서 중요한 요소로 작용하고 있습니다.

**2. 워드클라우드 분석**

두 번째 이미지에서는 주목할 만한 키워드들이 시각적으로 강조되어 나타나 있습니다. **아파트**와 **주택**이 가장 크게 나타나며, **사업**, **시장**, **상승** 등의 단어도 눈에 띕니다.

* **사업**: 부동산 개발 및 투자 사업이 활발히 진행되고 있음을 시사합니다. 이는 기업 및 정부의 부동산 정책과 연결될 수 있습니다.
* **상승**: 가격 상승에 대한 우려나 기대가 반영된 것으로 보입니다. 이는 시장의 불확실성과 투자자 심리를 반영합니다.
* **시장**: 부동산 시장의 변화와 관련된 다양한 요소가 있음을 나타내며, 정책 변화, 공급, 수요 등이 연관될 수 있습니다.

**인사이트**

1. **주거 형태의 선호**: 아파트와 주택에 대한 높은 관심은 주거 형태의 선호가 아파트 중심으로 이동하고 있음을 나타냅니다. 이는 안정성과 투자 가치가 높게 평가되기 때문입니다.
2. **서울 중심의 투자**: 서울 지역의 부동산 시장이 여전히 활발하며, 특히 아파트 분양에 대한 수요가 강한 상황입니다. 이는 수도권 인구 집중과 관련이 깊습니다.
3. **정책에 대한 반응**: 대출 규제나 정부 정책에 대한 반응으로서, 시장에서의 가격 상승 기대감이 반영되고 있습니다. 이는 투자자들에게 중요한 고려 사항이 될 것입니다.
4. **미래 전망**: 아파트 및 주택 시장의 지속적인 관심은 향후 부동산 시장의 안정성과 성장 가능성을 시사합니다. 기업과 정부의 정책 변화에 따라 시장의 방향성이 달라질 수 있으므로, 적극적인 모니터링이 필요합니다.

모델 결과



텍스트, 스크린샷, 직사각형, 도표이(가) 표시된 사진

자동 생성된 설명

### 고찰

이번 분석을 통해 부동산 현황을 파악하기 위해 빈도 분석과 감성 분석을 실시했습니다. 빈도 분석을 통해 특정 단어에 대한 관심도를 확인할 수 있었으며, 이는 부동산 시장의 주요 트렌드를 이해하는 데 유용했습니다. 특히, 아파트, 주택, 서울, 분양 등의 키워드가 높은 빈도로 나타나, 현재 시장에서 이들 요소가 중요하게 다뤄지고 있음을 알 수 있었습니다.

또한, 감성 분석을 통해 각 뉴스 기사에 대한 긍정적 또는 부정적 반응을 신속하게 확인할 수 있었습니다. 이를 통해 개별 기사를 일일이 분석하지 않고도 전체적인 시장 감성을 파악할 수 있어 효율적인 분석이 가능했습니다.

분석 기간은 2024년 10월 15일까지의 데이터를 수집하였으며, 이 시점에서 부동산 시장에 대한 반응은 여전히 긍정적인 것으로 판단되었습니다. 이는 향후 시장의 동향을 예측하는 데 중요한 지표가 될 것입니다.

하지만 T5 모델을 활용한 이진 분류 작업이 이루어지지 않은 점은 아쉬움으로 남습니다. 다양한 관점에서의 분석이 부족했던 점과 더불어, 모델의 활용을 통해 보다 심층적인 인사이트를 도출할 수 있었을 것이라는 가능성을 느꼈습니다.

한편, AutoML을 활용하여 성능이 우수할 것으로 예상되는 Lgbm모델을 선정한 결과, F1 score가 약 84%로 만족스러운 성능을 나타냈습니다. 이후 앙상블 기법이나 하이퍼파라미터 조정을 통해 성능을 더욱 향상시킬 수 있을 것으로 기대됩니다. 특히 긍정과 부정 데이터의 불균형 문제를 해결하기 위해 부정 데이터의 양을 늘린다면, 모델의 성능을 더욱 개선할 수 있을 것입니다.

결론적으로, 이번 분석은 부동산 현황을 이해하는 데 기여하였으며 향후 보다 다양한 분석 방법과 모델을 도입하여 심층적인 연구를 진행할 필요성을 느꼈습니다.