**한류의 세계화 요인 분석**

‘강남 스타일’의 텍스트 마이닝 결과를 바탕으로

텍스트이(가) 표시된 사진

자동 생성된 설명

**과목명. 데이터 마이닝**

**담 당. 손지은 교수님**

**제출일. 2022년 6월 10일**

**이름/학번. 조동혁 / 2016170864**

**김지후 / 2019170851**

**김우찬 / 2019390891**

**박석현 / 2019170345**



1. **서론**
   1. 한류 열풍의 중심 K-POP
   2. 딥러닝의 발전과 동떨어진 한류 연구 (문제 정의)
   3. 논문의 필요성과 방항셩
2. **본론**
   1. 데이터 선정 및 작업
      1. 데이터 선정과 이유
      2. 데이터 크롤링
      3. 데이터 전처리
   2. 데이터 분석
      1. 데이터 탐색
      2. 모델링
      3. 결과
   3. 결과에 따른 한류 성공 요인 분석
      1. 신나는 K-POP
      2. 한국 문화
      3. 대중적인 곡과 안무
3. **결론**
   1. 의의
   2. 한계
   3. 결론
4. **서론**
5. **한류 열풍의 중심 K-POP**

한류(Hallyu, Koran Wave)란 1990년대 중반 중국시장 진출을 계기로 동남아시아에서 시작하여 2000년대 초반부터 세계시장으로 퍼지기 시작한 한국 대중문화 인기현상을 지칭하는 용어이다. ‘다른 문화가 매섭게 파고 든다’ 는 의미인 ‘韓流’(한류)의 동음이어로서 1999년 중반부터 중국 언론에서 공식 사용하기 시작하였다.[1] 최근에 전세계에서 흥행한 영화 ‘기생충’, 넷플릭스의 ‘킹덤’, ‘오징어 게임’과 ‘지옥’ 그리고 현재도 계속 세계 신기록을 갈아치고 있는 BTS(방탄 소년단)을 통해 드라마/영화/K-POP 등 여러 분야에서 한류가 세계로 급속도로 퍼져 나가고 있음을 확인할 수 있다.

특히, K-POP이 신한류(新韓流)의 주역으로 급부상하면서 전 세계적으로 주목을 받고 있다. K-POP이란 대한민국에서 대중적 인기를 얻는 유행 음악을 총칭한 이름으로 간단히 가요, 또는 유행가라고 부르기도 한다. 영미권의 대중음악을 팝(POP, Popular Song)이라고 부르는데 기인해 영미권 이외의 국가들의 대중가요도 외국표기는 국가 이니셜을 붙여 나타내는 데서 기인하였고, 대한민국도 같은 관점에서 유래된 것으로 K-POP 이라고 표기한다. K-POP의 흥행은 싸이의 ‘강남 스타일’에서 본격적으로 시작되어 현재 BTS로 이어지며 지속된 상승세를 보이고 있다. K-POP은 유튜브나 페이스북, 트위터 등과 같은 다양한 소셜 네트워크 서비스를 통해 다양한 국가에 확산되고 있으며, 기억하기 쉽고 따라 부르기 쉬운 노래, 가수들의 세련된 스타일, 군무와 화려한 퍼포먼스 등을 통하여 전 세계의 소비자들을 매료시키고 있다.

1. **딥러닝의 발전과 동떨어진 한류 연구 (문제 정의)**

한류가 세계에 급속도로 퍼지고 있지만 아직 보완해야 할 부분도 많이 존재한다. 한류가 급부상함에 따라 한류를 더 거센 물결을 만들기 위해 많은 투자에 따라 수 많은 콘텐츠가 나오고 있지만, 흥행에 성공하는 비율은 아직 미미하다. 따라서, 한류에 대한 투자 증가에 따른 더 효과적이고 강력하게 퍼져나가기 위해 여러 콘텐츠 중 성공한 콘텐츠에 대한 연구를 통해 한류 돌풍의 원인 파악이 필요하다.

하지만 K-POP의 부상에 따른 연구는 아직 부족한 현실이며 연구들이 외국인들이 어느 부분에서 한류에 매력을 느꼈는지 실제 데이터를 가지고 분석한 연구는 더욱이 부족하다. 2000년대 들어서 강력한 컴퓨팅 성능에 따라 인공지능과 딥러닝이 여러 분야에서 강력하게 사용되고 있지만, 아직 한류에 대한 분석은 이에 발 맞추지 못하고 있다. 특히, 한류가 강력하게 퍼질 수 있게 한 SNS와 여러 OTT 플랫폼은 시간에 따라 내용이 급변하고 매우 유동적이며 확산되는 범위도 매우 광범위 하기에 인공지능을 통한 자동화된 분석이 더욱이 필요하다.

1. **논문의 필요성과 방향성**

딥러닝 기술을 사용한 한류에 분석이 부족함에 따라 이번 논문에서는 텍스트 마이닝(Text Mining) 기술을 이용해 실제 외국인들의 댓글을 분석해 보며, 외국인이 어느 부분에서 한류에 매력을 느꼈는지를 분석해 볼 것이다. 이로써 한류의 확산에 시행착오를 줄이고 더 빠르고 강하게 한류를 세계에 퍼트릴 수 있을 것으로 기대 된다.

본 연구는 기존의 연구들을 통한 문헌연구를 기초로 한류의 중심인 K-POP의 성공 사례를 분석하고 이를 바탕으로 K-POP을 통한 한류 활성화에 대한 구체적인 방안을 제시하고자 한다. 또한 인공지능과 딥러닝 기술을 한류 연구에 적용하는 기반을 마련할 수 있게 할 것이며, 본 연구는 탐색적 연구로 인하여 실증분석을 추후 실시하기 위한 기초적인 토대연구로 사용할 예정이다.

1. **본론**
2. **데이터 선정 및 작업**
   1. 데이터 선정과 이유

본 논문의 목적인 딥러닝을 통한 한류 성공 사례의 원인 분석은 싸이의 ‘강남 스타일’을 통해 하고자 한다. ‘강남 스타일’은 2012년 7월 15일 싸이(본명 박재상․36)의 정규 6집 ‘싸이 6갑(甲)’의 타이틀곡으로 코믹한 내용의 뮤직비디오와 독특한 구성의 ‘말춤’이 유튜브(Youtube)를 이용한 Viral marketing을 통해 전 세계적인 흥행돌풍을 일으킨 곡이다. 2013년 8월 5일 유투브 조회수 17억건을 돌파한 ‘강남스타일’ 은 유튜브 사상 최초로 10억 건을 돌파한 영상이자, 2022년 6월 기준 누적 조회수 44억회로 유튜브 조회순위 8위에 위치하고 있다.

한류의 신돌풍을 일으킨 곡이자 ‘유튜브 조회수’를 빌보드 차트에 반영하게 만든 곡인 강남스타일은 신한류의 시작점이라고 봐도 무방하다. ‘강남 스타일’은 특히 유튜브에서 큰 히트를 쳤고 해당 뮤비엔 500만개 이상의 댓글이 달려 있다. 따라서 유튜브의 ‘강남 스타일’ 뮤비의 댓글을 텍스트 마이닝을 통해 분석해 보면 외국인들이 K-POP과 더 나아가 한류의 어느 부분에서 매력을 느꼈는지 알 수 있을 것이다.

* 1. 데이터 크롤링

|  |
| --- |
| 텍스트이(가) 표시된 사진  자동 생성된 설명 |
|  |

(라이브러리 호출 코드 및 Queue Class 생성 코드)

유튜브에 업로드 되어 있는 싸이의 강남스타일 영상의 댓글을 크롤링하기 위해 pandas와 및 다양한 라이브러리들을 불러 들인다. 이중 Selenium은 다양한 언어에서 웹 드라이버를 통한 웹 자동화 테스트를 지원하는 라이브러리로 웹에서 자동화를 통한 다양한 작업을 가능케 한다.

이번 프로젝트에서는 기존에 웹 크롤링시 많이 사용하는 Scrolldown 방식이 아닌 [Queue]를 이용한 방식을 사용했다. 기존의 Scrolldown 방식은 크롤링을 하기 위해 먼저 스크롤을 페이지 제일 밑까지 내려야 하는데 이 경우 스크롤을 내릴 때 마다 약 3초 간의 Loading시간이 존재해 댓글 수가 많은 경우 크롤링 시간이 길어질 수 밖에 없다.

하지만 [Queue] 방식을 이용할 경우 1. Queue 생성하기 2. 0.1초마다 ScrollHeight(컨텐츠 전체 높이)를 Engueue하기 3. Enqueue시 Queue의 첫번째 데이터와 비교하여 다를 경우 Scroll이 업데이트 된 것을 의미하므로 새로운 ScrollHeight 앞의 데이터를 모두 Dequeue하기 4. Queue가 꽉 차는 경우(0.1 sec \* Max queue size의 시간이 흘러도 Scroll Height가 업데이트 되지 않는 경우) 마지막 Scroll까지 내린 것으로 간주하고 댓글 크롤링을 시작한다. 또한, Max Queue Size를 조절하여 스크롤을 모두 내렸다고 판단할 수 있는 시간을 조절할 수 있어, 스크롤을 내리다 중간에 멈추는 현상을 개선할 수 있다.

|  |
| --- |
| 텍스트이(가) 표시된 사진  자동 생성된 설명 |
|  |
| 텍스트이(가) 표시된 사진  자동 생성된 설명 |

Selenium의 성능을 향상시키기 위해 옵션을 설정하고 크롬드라이버 창크기/그래픽성능 낮춰서 크롤링 성능 높이기 위한 네트워크 및 사이트 설정 코드이다.

Diver.get으로 크롤링 목표(사이트) 설정한 뒤 Scroll Down과 제한시간(50초) 설정하고 Scroll Down이 빠르게 될 수 있도록 Scroll Height를 가져오는 주기를 0.1초로 설정한다. Enqueue시 Queue의 첫번째 데이터와 비교하여 다를 경우 Scroll이 업데이트 된 것을 의미하므로 새로운 ScrollHeight 앞의 데이터를 모두 Dequeue 한다.

|  |
| --- |
|  |

for문을 만들어 크롤링에 필요한 데이터(댓글 내용, 유튜브ID, 댓글 날짜, 좋아요 개수)를 find 함수로 찾아 가공할 수 있도록 설정하는데 데이터를 수집하는 과정에서부터 깔끔하게 전처리할 수 있도록 가공하는 부분까지 함께 작성했다. 마지막으로 완성된 데이터 리스트를 데이터 프레임화하여 엑셀로 만들었다.

* 1. 데이터 전처리

|  |
| --- |
|  |
|  |
|  |

Spacy는 파이썬에서 자연어처리기반 Opensourse library이다. Spacy에서는 Tokenization, 품사태깅, 개체인식 등 기본적인 전처리 기능을 지원해주고 있어서 이번 프로젝트에 사용했다.

|  |
| --- |
|  |

크롤링 결과로 약 45,000개의 Youtube 댓글을 가져와서 전처리진행을 해주었다. 우선 형태를 List로 변환했고 댓글들 중 특수 형태들: Emails, new line characters, distracting single quotes 등을 제거했다.

1. **데이터 분석 및 모델링**
   1. 데이터 탐색

|  |
| --- |
| 텍스트이(가) 표시된 사진  자동 생성된 설명 |
|  |

각 문장들을 단어들의 list형태로 토큰화 해주었고, 댓글들의 punctuations을 제거 해주었다. Stemming은 word를 word stem (원형)으로 변환을 해주는 기능입니다. 그리고 명사, 형용사, 동사, 부사를 Lemmatization해주었습니다. 이렇게 함으로써 딕셔너리 안에 존재하는 특수단어들의 수를 줄일수가 있었습니다.

* 1. 데이터 모델링

|  |
| --- |
|  |
|  |

CountVectorizer를 사용해서, LDA 토픽모델 알고리즘을 구축했다. 추출할 총 토픽의 개수는 20개로 우선 설정을 해주었으며, 추후에 Grid search를 사용해서 변경을 해줄 예정이다. Log-Likelihood 및 perplexity를 찾아주었고 Log-Likelihood는 높고 Perplexity는 낮을 수록 좋은 모델이다.

|  |
| --- |
|  |

이 부분에서는 Grid Search를 사용해서 최적의 LDA model를 구축하려 했습니다. 토픽의 수를 결정하려고 했으며, Learning rate를 어떤게 control 하는지 또한 하려고 했습니다.

여기서 Grid Search란, “Grid search (격자 탐색) 은 모델 하이퍼 파라미터에 넣을 수 있는 값들을 순차적으로 입력한뒤에 가장 높은 성능을 보이는 하이퍼 파라미터들을 찾는 탐색 방법이다.” 쉽게 설명을 하자면, 모델이 학생이라고 가정을 할 때, 학생이 공부를 하는 여러가지 (학습모델)들이 존재한다, 이때 각 학습모델의 세부적인 규율들을 따진 후, 어떤 방법이 제일 적합한지를 찾는 것이 Gid Search라고 합니다.

|  |
| --- |
|  |

단어가 특정한 Topic에 알맞은지 확인하기 위해서는 어떤 단어가 가장 기여도가 높은지를 파악하면 됩니다. 이 점을 활용하기 위해 위에 처럼 Best component 10을 근거로 10개의 Topic을 만들고 각 토픽과 데이터(Doc)를 이용해 Dataframe을 만들었습니다. 이때 녹색으로 하이라이트가 된 부분은 각 column에서 Dominant한 토픽입니다. (0.1보다 높은 Value는 녹색으로 구분)

|  |
| --- |
|  |

Topic – Keyword Matrix를 만들고 Vectorizer를 통해 각 Value를 계산했다.

* 1. 데이터 결과

|  |
| --- |
|  |

Topic Labeling을 위해 총 10개의 Topic을 만들었고, 가장 높은 Value를 가진 word 15개[어떻게 value을 측정했는지]를 선정하여 Dataframe으로 다시 구축했다.

|  |
| --- |
|  |

각 Topic별 word를 상의 후 알맞은 토픽을 선정해 주었습니다. 모든 단어들을 고려하기보단, 중요하고, 겹치는 단어 4~5개 위주로 선정을 해주었습니다. 결과는: Style, Childhood, Music, Fun, Diverse Culture, Nostalgia, Exciting, 실력, 남녀노소, SNS로 구별을 해주었습니다.

1. **분석 결과**

위의 Top Topic을 통한 ‘강남 스타일’의 성공 요인은 총 3개 분류로 나눌 수 있다. 첫째로는 ‘Music’, ‘Fun’, ‘Exciting’과 ‘실력’을 포함하는 신나고 재미있는 완성도 있는 음악, 다음으로는 ‘Diverse Culture’을 통한 한국의 문화에 관한 부분과 마지막으로 ‘Nostalgia’, ‘남녀노소’와 ‘SNS’를 묶어 대중적으로 인기를 얻을 수 있는 음원과 춤으로 나눌 수 있다.

* 1. Exciting한 요소

자연, 불, 사람, 실외이(가) 표시된 사진

자동 생성된 설명   
(왼: ‘싸이’의 ‘강남스타일’ MV / 오: ‘BTS’의 ‘DNA’ 앨범)

먼저 ‘Music’, ‘Fun’, ‘Exciting’과 ‘실력’이 Top Topic에 있는 것을 통해 외국인들이 신나고 경쾌한 K-POP에 매력을 느낌을 알 수 있다. ‘강남 스타일’은 신나고 흥을 고조시키는 음악과효과들이 들어가 있고 이는 강남스타일에 대해 가장 많이 말한 내용이다. 실제로 K-POP의 곡들은 발라드와 같이 서정적인 곡들보다 팝 종류의 경쾌한 곡들이 많다. 세계적인 돌풍을 일으키고 있는 BTS의 대다수의 곡들도 경쾌한 박자를 가진 곡이 많다.

* 1. 한국 문화

   
(왼: ‘강남스타일’ MV에 등장한 고속버스 / 오: ‘NETFLIX’의 ‘오징어 게임’)

Top Topic에 ‘Diverse Culture’가 있음을 통해 다양한 한국의 문화가 해외에서 좋게 작용함을 알 수 있다. ‘강남 스타일’에는 ‘놀이터’, ‘사우나’, ‘고속버스 속 노래’와 ‘한강 조깅’ 등 다양한 한국 문화가 등장하였고 외국인들은 이를 흥미롭게 생각했다.

실제로 한류 드라마인 ‘오징어 게임’은 한국 고유의 ‘딱지 치기’, ‘달고나 게임’, ‘무궁화 꽃이 피었습니다.’와 ‘오징어 게임’이 등장하는데, 외국인들은 이에 대한 호기심과 흥미를 보였고 실제로 많이들 따라했다. 또한, 사극인 ‘킹덤’에서 나오는 한국 전통 모자인 ‘갓’에도 외국인들은 흥미를 보였다.

* 1. 대중적일 수 있는 음악과 안무

 

(좌: 미국에서 ‘강남 스타일’의 말춤을 선보이는 ‘싸이’ / 우: 해외 플래시게임으로 제작된 '강남스타일')

마지막으로 ‘남녀노소’와 ‘SNS’를 통해 대중적인 음악과 춤이 ‘강남 스타일’의 확산에 좋은 영향을 줌을 확인할 수 있다. 실제로 ‘강남 스타일’의 경우 해외에서 많은 ‘밈’(Meme)이 SNS를 통해 퍼지고 여러 셀럽이 ‘강남 스타일’을 SNS에 올리며 확산되었다. 그리고 남녀노소 누구나 출 수 있는 춤사위를 통해 강남 스타일은 대중화를 이룰 수 있었다. 또한, 이렇게 대중화를 한 번 이루게 되면 많은 사람들이 공통적으로 이에 대한 ‘Nostalgia’(향수)를 가지고 있어 시간이 지나도 계속 찾아 보게 만들었다.

1. **결론**
2. **연구의 의의**

이번 연구를 통해 딥러닝을 한류 연구에 적용하는데 초기 모델을 만든 것과 유튜브의 댓글을 크롤링해 데이터 셋으로 씀으로써 더 정확하고 현실적인 분석을 가능케 한 것에 의의가 있다. 이 분석 결과에 따라 앞으로 K-POP은 신나고 경쾌한 멜로디와 대중적인 안무를 중심으로 발전시켜야 함과 한류 전반적으론 한국 고유의 문화를 가감없이 보여주며 외국인들이 이에 더 흥미와 관심을 갖을 수 있게 해야 함을 알 수 있다.

1. **연구의 한계**

이번 연구의 한계는 영어 자막만을 분석하고 나머지 언어의 댓글들에 대한 분석은 이뤄지지 못한 것은 이번 연구의 한계이다. 추후에 언어 번역기능도 사용해 여러 언어로 쓰여진 댓글도 분석하면 더 정확도가 올라갈 것이다. 또한, 전처리 과정에서, 영어를 제외한 언어들을 제거할때, 영어랑 같은 알파벳을 사용하고 있는 언어들을 100% 제거하지는 못했던거 같다, 특히 스페인어 같은 경우는 영어랑 공통점이 많이 있기에, 제거를 완벽히 하기에는 힘들었던거 같다. 다음에 비슷한 프로젝트롤 진행 할시에는, 스페인어 딕셔너리를 따로 적용후, 그 부분만 제거를 하는 형식으로 진행을 해주면 해결이 될거 같다.

또한 한류를 좋게 보는 시선만 분석하고 반대로 ‘혐한’ 등과 같이 한류에 반감을 가지고 있는 사람들의 의견을 분석하지 못한 것도 한계이다. 추후에 한류에 대한 악평에 대한 분석도 같이 한다면, 이를 통해 한류의 강점을 더 키우고 약점을 보완하며 한류를 더 효과적으로 발전시킬 수 있을 것이다.

LDA과정에서, 마지막에 주어진 단어들을 보고 알맞은 토픽들을 팀원들과 함께 상의후 정해주었다. 하지만, 이 과정에서 모든 15개의 단어를 활용해서 알맞은 토픽을 정하는데는 한계점이 있었다. 그러므로, 15개의 단어를 모두 고려하기 보다는, 그중에서 중요하다고 생각이 드는 5~8개의 단어들을 가지고 토픽을 만들었다.

크롤링을 하는 과정에서는 싸이의 <강남스타일[[1]](#footnote-1)> 영상에 있는 댓글만 크롤링을 진행 해주었다. 다음에 진행 할 시에는 이 영상 뿐만 아니라, 홍보 영상, 다른 유튜버의 영상등에서 코멘트를 같이 크롤링 해오면, 조금 더 정확한 분석이 가능할거 같다.

1. **결론**

싸이의 ‘강남스타일’을 텍스트 마이닝으로 분석을 해보고 다른 성공한 한류 콘텐츠와도 비교해보며 분석을 더 구체화 할 수 있었다. 한류에 대한 올바른 이해를 바탕으로 올바른 방향성을 설정할 수 있게 되었다.

위에 언급 했듯이, 강남스타일의 성공요인을 다양한 이유가 있겠지만, 이번 프로젝트를 통해서는 Music, Fun, Exciting 세가지 요소로 정리를 해봤습니다. 사람들은 대부분 댓글을 통해 얼마나 노래가 신나고 Active 한지 표출을 많이 해주었으며, 음악적인 부분에서 느껴보지못한 다른문화인 한국문화를 느끼면서 호기심을 보여주었습니다. 시간이 지나도, 그들이 처음 들었을 때 느꼈던 신선함과, Fun한 춤이 강렬하게 기억속에 남아있었으며, 그러한 내용들을 댓글을 통해서 자주 보여주는 모습을 보여주었습니다.

또한, 춤 같은 경우도, 그들이 알고있는 단조롭거나 멋있는 (또는 섹시한) 춤 보다는 짧지만 굵은 “말춤”에도 굉장한 반응을 보였습니다. 아무래도 통상적으로 어떤 노래의 안무는 어렵고 대중들이 쉽게 따라할 수는 없는 춤이 대부분이지만, 강남스타일의 “말춤”은 누구나 영상을 한번 보면 쉽게 접근이 가능한 춤이기 때문인거 같습니다. 이러한 세가지의 조화가 잘 이루어지면서, 강남스타일은 대중들에게 큰 매력요소로 다가왔을 것이며, 외국에서 “한국”이라는 나라를 대표하는 많은 것들중 하나로 자리매김 된거 같습니다.

**참조**

[1] 송기문, 신성아, and 서헌. "K-POP 을 통한 한류관광 활성화 방안: 싸이 (PSY) 의 강남스타일을 계기로." 관광레저연구 26.3 (2014): 63-76.

[2] 조인희, and 윤여광. "한류문화에 영향을 미친 Youtube 파급효과에 관한 연구: 가수 싸이 (PSY) 신드롬을 중심으로." 한국엔터테인먼트산업학회논문지 7.2 (2013): 9-18.

[3] 양문희. "한류 방송 콘텐츠 확산을 위한 방안 연구." 한국콘텐츠학회논문지 19.11 (2019): 201-210.

[4]

[5]

1. <https://www.youtube.com/watch?v=9bZkp7q19f0&ab_channel=officialpsy> [↑](#footnote-ref-1)