산업통상자원부 공공데이터 활용 아이디어 공모전 분석 결과 제출 양식

1 명칭

ㅇ해외 바이어와 국내 기업 텍스트 기반 매칭 모델

2 제안배경

벤처기업협회와 KOTRA가 공동으로 벤처·스타트업 271개 사를 대상으로 진행한 설문조사에 따르면, 해외 진출을 시도하는 벤처 기업들이 현지 정보 부족, 검증된 바이어 발굴 실패 등을 이유로 바이어 매칭에 큰 어려움을 겪고 있는 것으로 나타났다. 기존에는 이러한 문제를 해결하기 위해 컨설팅 업체나 벤더 사를 통해 수출 과정을 진행했지만, 이 방식은 명확한 문제점을 가지고 있다.

첫째, 바이어의 흥미를 유발하고 니즈를 파악하는 데 상당한 시간이 소요된다. 이는 바이어가 관심을 가질 만한 제품을 토대로 제조사의 역사와 성과를 살펴 바이어와의 매칭 가능성을 상상해야 하기 때문이다.

둘째, 매칭이 성사되지 않을 경우, 각 당사자(바이어와 국내 기업)에게 시간 낭비를 초래한다. 또한, 중간에 개입한 업체도 난감한 상황에 처하게 된다. 거래 과정에 처한 모두가 요구와 기대를 조율하기 위해 많은 자원을 투입하지만, 매칭이 실패할 경우 투자한 것들에 대한 보상을 받을 수 없는 것은 결국 모두에게 부담으로 이어진다.

100개 업체를 접촉하면 약 5개 업체와만 거래 조율이 이루어지는 결과는 매칭 과정의 비효율성을 잘 보여주며, 이는 곧 기업의 해외 진출 전략에 있어서 큰 장애물이 된다.

이와 같은 문제를 해결하기 위해 KOTRA는 'AI 활용, 실거래 데이터 기반의 수입확률 분석·매칭 지원사업'을 진행하였고, 이는 국내 기업에 맞는 신규 바이어 발굴과 수출 성약률을 높이는 성과로 이어졌다. 다양한 바이어와 기업 간 수출 계약이 체결되면서, 더 간편하고 높은 확률을 가진 매칭 모델의 중요성을 입증했다.

기존의 컨설팅 업체나 벤더 사를 통한 구조를 개선하여, 수요와 공급의 만남 지점을 정확히 예측할 수 있는 모델을 활용하는 것은 단순히 바이어와 셀러를 매칭하는 것을 넘어 바이어의 구매 확률을 높이는 데 기여한다. 이를 기반으로, 매칭 모델을 도입하여 효율적이고 정확한 매칭 과정을 구축할 필요성이 대두되고 있다.

이에 따라, 적합한 바이어를 보다 효과적으로 발굴하는 것을 목표로 하며, 상품을 잘 아는 셀러와 시장을 잘 아는 바이어 중간에서 이를 가장 잘 이해할 수 있는 매칭 모델을 제안하고자 한다.

3 분석 내용 및 분석 결과

[HSCODE]

HSCODE란 국가 간 원활한 상품 거래를 위해 상품의 종류를 숫자 코드로 분류한 것이다. 이 코드의 앞 6자리는 국제 공통으로 사용되며, 2자리마다 류(앞 2자리)/호(앞 4자리)/소호(앞 6자리)로 나뉜다. 대한민국에서는 HSCODE에 HSK 4자리를 추가하여 세부 분류를 진행하고 있다. HSCODE의 2자리 단위마다의 의미를 활용하여, 2/4/6/10 단위 순서로 매핑 범위를 점점 좁혀 나가는 방식으로 분석을 진행하였다.

[데이터 분석]

①통계청 국제표준산업분류 HSCODE 6단위 매핑

- [ISIC4]: 375종류, [KSIC10]: 996종류, [HS2017]: 5359종류
- [HS2017]이 한 개 이상 매핑되어 있는 [ISIC4]: 185종류
- 185종류의 [ISIC4]는 각 평균 30.18개의 [HS2017] 6단위와 매핑되어 있다.

②비식별된 해외기업별 영문 텍스트데이터

- 비식별된 해외기업별 영문 텍스트데이터의 ISIC4 코드([CODE]) 10,000개 중 약 3,000개는 한 개 이상의 [HS2017]과 매핑되어 있다.(통계청 국제표준산업 분류 HSCODE6단위 매핑을 기준으로 한다.)

③관세청_HS부호

- HSK에 대한 [영문품목명]을 HSK에 대한 설명으로 사용한다. [영문품목명]이 Other인 경우에는 매핑된 성질통합분류코드명을 영어로 번역하여 추가한다.

④UNSD-HS2017

- HS2017을 류/호/소호 단위로 설명하는 영문 텍스트 데이터를 포함하고 있다.

[모델 선택]

본 분석은 'BERT 모델'을 채택하여 데이터 학습 및 예측을 진행하고 있다.

1. 영문 텍스트의 문맥 이해

- 단순 단어 나열이 아닌 문장 형태로 제공되는 데이터를 처리하기 위해, 문맥을 효과적으로 파악할 수 있는 모델이 요구된다. BERT는 문맥을 '양방향'으로 이해하여 문맥을 정확히 파악할 수 있기 때문에, 문장 이해에 강점이 있는 BERT를 채택하였다.

2. 텍스트 문장의 복잡성

- 문장 길이가 긴 기업 DSC를 처리하기 위해, RNN 계열의 모델보다 데이터 내의 관계를 추적해 맥락과 의미를 학습할 수 있는 트랜스포머 기반 모델 BERT를 채택하였다.

3. 모델의 확장성 및 정확도

- BERT는 대량의 데이터로 사전 학습된 모델로써, 소량의 데이터로도 전이 학습을 통해 뛰어난 성능을 보여준다. 따라서, 제한된 데이터를 활용해야 하는 해당 과제에서 목적에 맞는 학습을 수행하는 데 적합할 것이라 판단하였다.

[알고리즘]

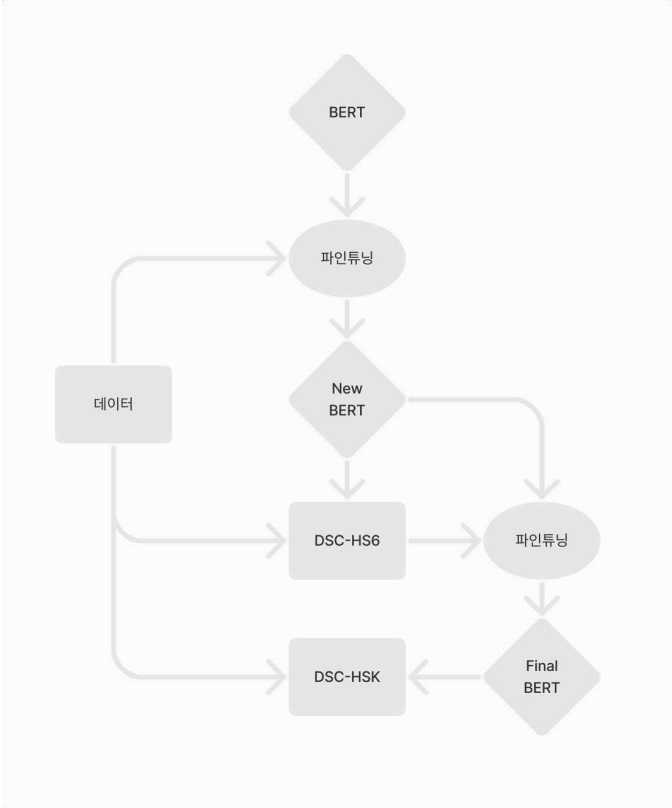

그림 1 알고리즘 순서도

BERT 파인튜닝 -> DSC - HSCODE 6단위 매핑 -> BERT 추가 학습 -> DSC - HSK 매핑의 순서로 이루어진다.

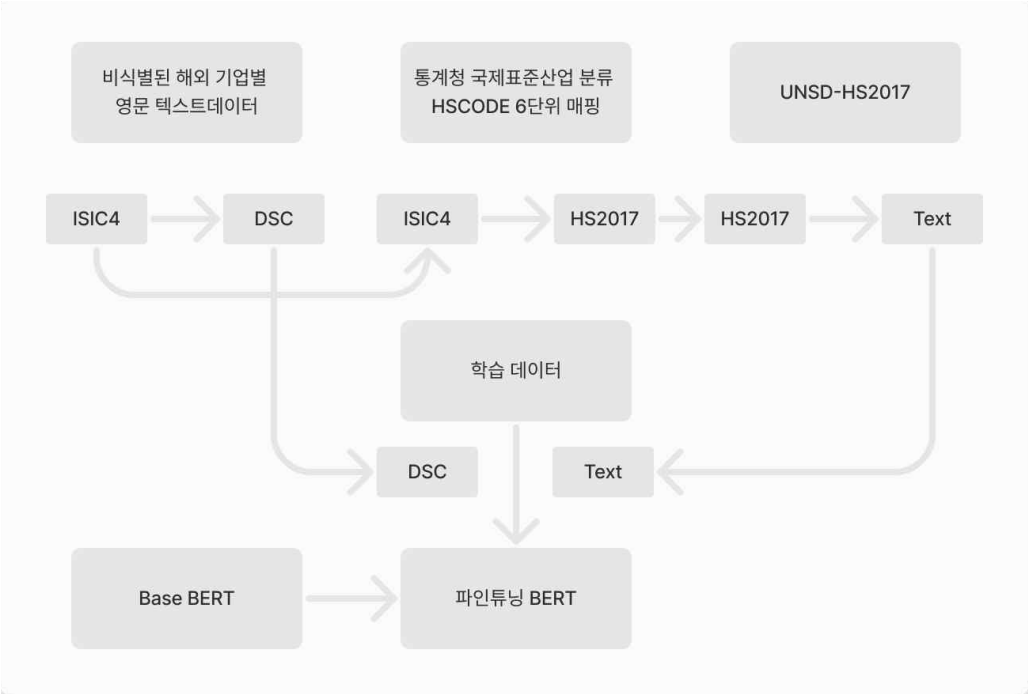

그림 2 BERT 파인튜닝 과정에 대한 그림

<BERT 파인튜닝>

(3페이지의 그림2를 기반으로 한다.)

1. 비식별된 해외기업별 영문 텍스트데이터에서 DSC와 ISIC4를 가져온다.
   -> [DSC, ISIC4] 10,000개
2. 통계청 국제표준산업분류 HSCODE 6단위 매핑을 통해 얻은(ISIC4 -> HS2017) 데이터를 활용하여, ISIC4를 HS2017로 변환한다. 이때, ISIC4에 매핑된 HS2017이 없을 경우 사용하지 않는다.
   -> [DSC, HS2017]
3. HS2017의 설명을 도울 외부데이터(UNSD-HS2017)를 활용하여 HS2017을 Text로 변환한다.
   -> [DSC, Text]
4. 추출된 문장 쌍 데이터를 Label = 1로 적용하여, BERT 학습데이터로 사용한다.

<BERT 파인튜닝>의 목적: DSC와 HS2017의 Text 문장 쌍 학습을 통해, 모델이 DSC와 Text 간의 유사성을 정확하게 판단할 수 있다. 학습된 모델을 사용하여 매핑된 HS2017이 없는 DSC에 대하여 HS2017의 매핑을 명확하게 수행하고자 한다.

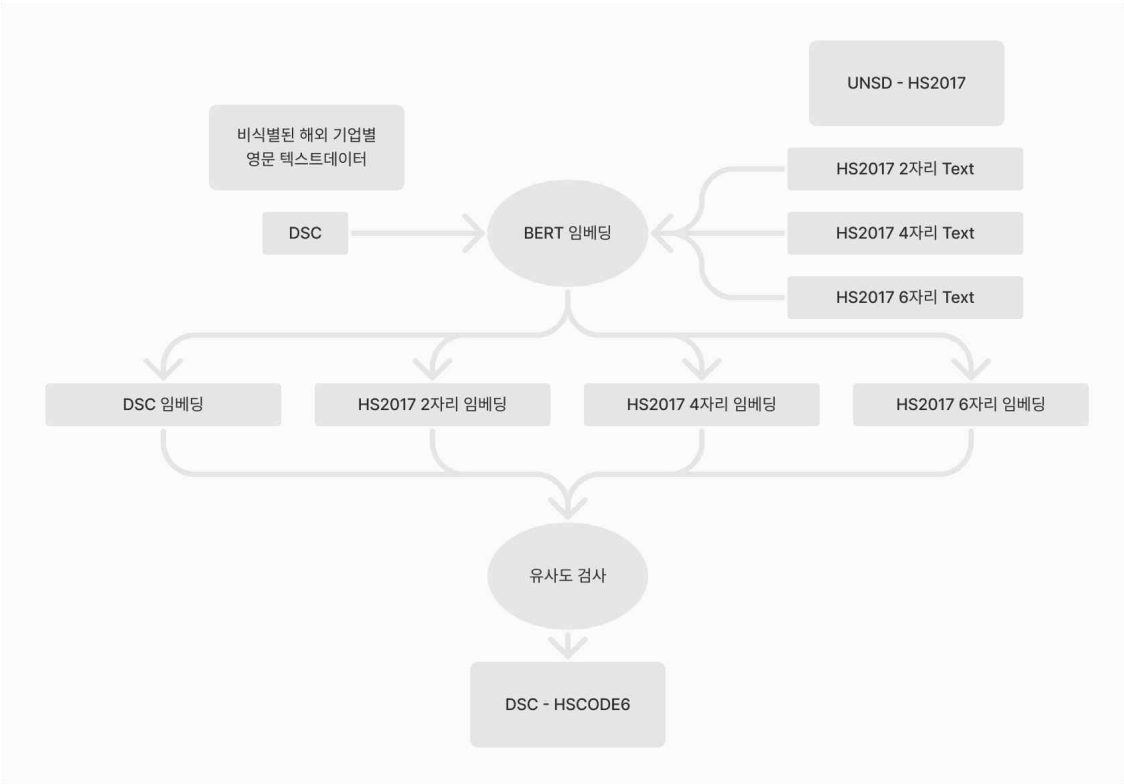

그림 3 DSC - HSCODE 6단위 매핑 과정에 대한 그림

<DSC – HSCODE 6단위 매핑>
(4페이지의 그림3을 기반으로 한다.)

1. 파인튜닝된 모델을 활용하여 DSC와 HS2017_Text를 임베딩한다.
   -> DSC_emd: 10,000개, HS_2자리_ebd: 97개, HS_4자리_ebd: 1,223개, HS_6자리_ebd: 5,388개
2. DSC_ebd와 HS_2자리_ebd를 '코사인 유사도'를 활용하여 추출한다. 이때, 추출 기준은 max 코사인 유사도 값의 85%이상으로 한다.
   -> [DSC, HScode_2]
3. DSC_ebd와 HS_4자리_ebd를 '코사인 유사도'를 활용하여 추출한다. 이때, 추출 기준은 max 코사인 유사도 값의 85%이상으로 하며, HS_4자리는 추출된 HS_2자리의 하위 코드만 사용한다.
   -> [DSC, HScode_4]
4. DSC_ebd와 HS_6자리_ebd를 '코사인 유사도'를 활용하여 추출한다. 이때, 추출 기준은 max 코사인 유사도 값의 85%이상으로 하며, HS_6자리는 추출된 HS_4자리의 하위 코드만 사용한다.
   -> [DSC, HScode_6]

<DSC – HSCODE 6단위 매핑>의 목적: 통계청 HSCODE 6단위 매핑 데이터만으로는 HS2017과 연결할 수 없던 7,000개 가량의 DSC에 대해 HSCODE 6단위 매핑을 하고자 한다. 이때, 코사인 유사도 최댓값을 기준으로 추출하여 최소 1개의 HSCODE 매핑을 보장하였으며, 높은 유사도를 보이는 HSCODE가 많을 경우 추출되는 HSCODE의 수도 늘어나도록 구현하였다.
-> DSC 1개 당 평균 45개의 HSCODE가 매핑되었다. 이는 기존 30.18개와 유사한 값이다.

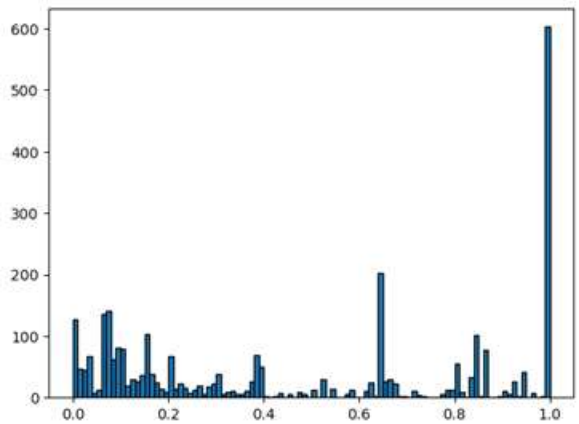

그림 4 X축-정밀도, Y축-개수에 대한 그래프

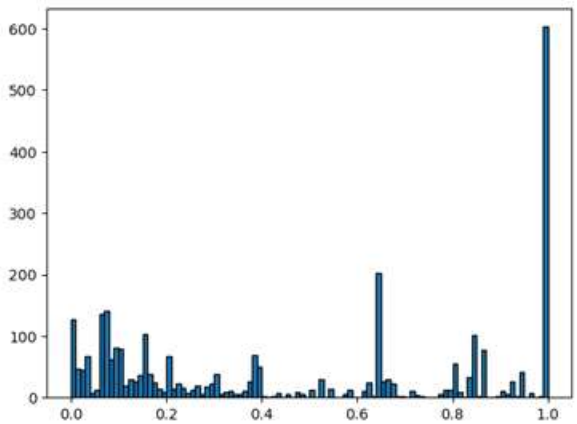

그림 5 X축-재현율, Y축-개수에 대한 그래프

위의 그림4, 그림5를 참고하였을 때, 통계청 HSCODE 6단위 매핑과 비교하였을 때 유사한 결과를 보이는 것을 확인할 수 있다.

<BERT 추가 학습>

1. 위 <DSC – HSCODE 6단위 매핑>의 결과물인 [DSC, HScode_6] 10,000개 중 기존에 매핑이 되어있지 않던 7,000개를 추출한다.
   -> [DSC, HScode_6] 7,000개
2. 위 <BERT 파인튜닝> 과정을 적용하여 BERT 추가 학습을 진행한다.

학습에 이용되지 않은 7,000개의 DSC를 학습에 포함시킴으로써 HSK 매핑 단계에서의 정확도 향상을 목적으로 학습하였다. Final BERT와 New BERT를 이용하여(각 용어는 '그림1 알고리즘 순서도' 참고) DSC-HSK 매핑을 진행해 보았을 때, 7,000개의 DSC를 추가 학습 시킨 3단계의 BERT가 더 좋은 성능을 보여주는 것으로 확인되었다. (이때, 성능 평가 기준은 한국무역통계진흥원의 HSCODE 내비게이션을 활용하여 비교하였다.)

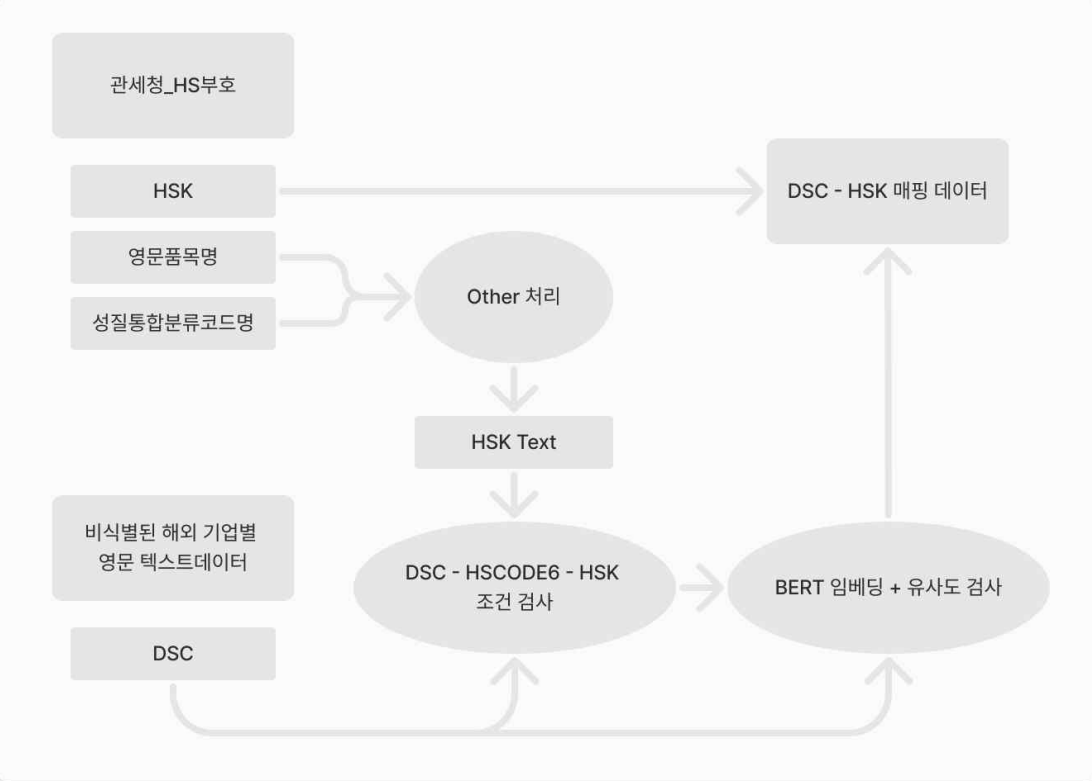

그림 6 DSC – HSK 매핑 과정에 대한 그림

<DSC – HSK 매핑>

1. 관세청_HS부호의 영문품목명을 HSK의 설명으로 사용한다. 이때, 영문품목명이 Other일 경우 성질통합분류코드명을 번역하여 사용한다.
   -> [HSK, HSK_Text]: 12,422개
2. HSK_Text와 DSC를 추가학습한 BERT로 임베딩한다.
   -> DSC_ebd: 10,000개, HSK_영문품목명_ebd: 11,294개
3. <BERT 추가 학습> 과정에서 얻은 [DSC, HScode_6]을 활용하여, 매핑된 HScode6의 하위 HSK와 DSC를 기반으로 코사인 유사도를 추출한다. 이때, 추출 기준은 최댓값의 80%로 하며, 유사도를 기반으로 최종적으로 매칭된 HSK를 출력한다.

<DSC - HSK 매핑>

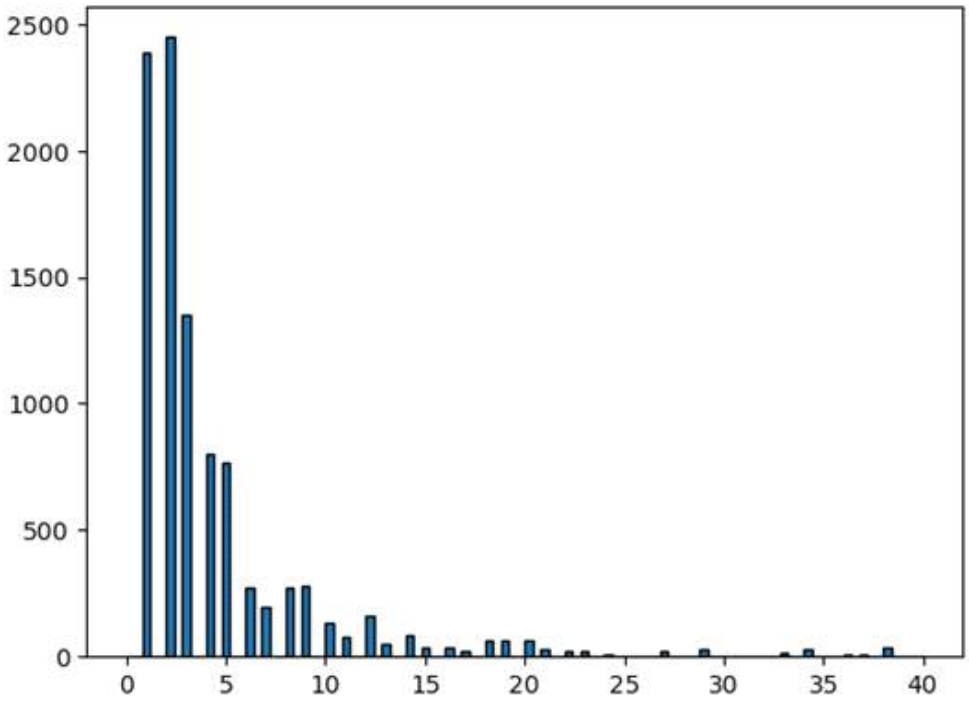

그림 7 X축-기업 ID 당 매핑된 HSK 수, Y축-X축에 해당하는 ID 개수에 대한 그래프

위의 그림7을 참고하였을 때, 대부분의 기업 ID에 1개에서 5개 사이의 HSK 코드가 매핑된 것을 확인할 수 있다.

4 활용데이터

①비식별된 해외기업별 영문 텍스트데이터
- [DSC]를 임베딩하여 모델 학습과 예측에 적극 사용한다.
- [CODE]와 '통계청 국제표준산업분류 HSCODE 6단위 매핑 데이터'와의 연결을 활용한다.

②관세청_HS부호_240101
- [HS부호]는 최종적으로 매핑하고자 하는 HSK로 사용한다.
- [영문품목명]은 HSK의 설명 텍스트로써 학습에 사용한다.
- [성질통합분류코드명]은 [영문품목명]의 부족함(예. [영문품목명]이 Other인 경우)을 보충하기 위해 사용하며, 영어로 번역하여 사용한다.

③통계청 국제표준산업분류 HSCODE 6단위 매핑
- 매핑되어있는 [ISIC4]와 [HS2017]을 모델 학습과 예측에 적극 사용한다.

④UNSD-HS2017(https://comtradeapi.un.org/files/v1/app/reference/H5.json)
- HS2017을 영문 텍스트로 설명하는 데이터이다.
- 류/호/소호로 나누어서 설명하는 데이터이다.
- 별도의 번역 과정 없이 '비식별된 해외기업별 영문 텍스트데이터'의 [DSC]와의 정확도를 평가할 수 있다.
- HS2017 코드의 의미와 세부 내용을 보다 명확하게 이해하는 데 도움을 주는 데이터로써, 모델이 보다 정확하게 학습하고 예측할 수 있도록 돕는다.

5 사업화방안 및 기대효과

[서비스의 시장성 및 상용화 가능성]

중소벤처기업진흥공단에서 진행한 설문에 따르면, 응답자의 98.2%가 글로벌 시장에 진출할 계획이 있다고 답변했다. 그러나 많은 기업들이 바이어 및 파트너 발굴의 어려움(34.6%), 해외시장 정보 부족(33.4%)으로 인해 어려움을 겪고 있는 것으로 나타났다. 이러한 문제를 해결할 수 있는 모델에 대한 수요는 매우 높을 것이며, 특히 검증된 바이어 발굴에 대한 필요성이 절실한 상황이다.

서비스의 주 고객층은 '효과적인 잠재 바이어와의 매칭 과정을 원하는 국내 기업', 이에 따른 서브 고객층은 '니즈에 맞는 제품을 수출하는 기업을 원하는 해외 바이어'로 설정한다. 확보된 타켓층을 바탕으로 KOTRA와 협력하여, 국내 벤처 및 중소기업을 대상으로 시범 운영하고자 한다. 이후 지속적인 거래 데이터 업데이트를 통해 모델의 정확도를 향상시킬 계획이다. 한국 시장에서의 성과를 글로벌 시장으로 확장하기 위해서는 적합한 해외 바이어를 효과적으로 발굴하는 작업이 필수적인 만큼, 보다 정확한 매칭 모델의 수요는 계속 늘어날 것이다.

[기대효과]

(국내 기업)검증된 바이어와의 매칭을 통해 신규 시장 개척 기회를 얻음.
(해외 바이어)정밀하게 분석된 니즈를 바탕으로 기업을 매칭 받을 수 있음.

국내 기업들의 원활한 매칭으로 인한 수출 증대는 국가 경제에 기여할 수 있을 것이며, 바이어 발굴을 모델에게 맡김으로써 제조사는 절약된 시간과 비용을 제품에 투자하여 제품 개발에 더욱 몰두할 수 있을 것이다.

기존의 비효율적인 매칭 프로세스를 개선하는 것은 어려움을 극복하는 데 중요한 역할을 한다. 중간 업체를 통해 이루어지던 바이어와 셀러 간의 매칭 과정은 상담 및 견적 요청부터 시작하여 시장 조사, 바이어와의 개별 연락, 그리고 매칭 결과를 기다리는 일련의 절차를 포함한다. 이 과정은 매우 복잡하고, 서비스 시장 조사를 진행하는 데만 최소 1개월 이상이 소요되어 기업들에게 상당한 부담을 준다. 이 모델은 많은 인적자원, 시간, 수천만 원에 달하던 비용을 절감하는 데 큰 기여를 할 것이며, 이 모델의 활발한 활용을 통해 국내 제조품들이 다양한 산업과 국가로의 수출로 이어질 수 있기를 기대한다.