반려동물 질병상담 챗봇 서비스 구현

배주현[1], 성예원[2], 육예은[3], 장윤희[4]
[1]서울과학기술대학교 컴퓨터공학과
[2]배화여자대학교 스마트IT학과
[3]강남대학교 사회사업학과
[4]상명대학교 경제금융학부
qowngus33@gmail.com, sungyaeone@gmail.com,
yy.eun6@gmail.com, lightson23@naver.com

Chatbot for Diagnosis of Pet diseases : Service Development and Distribution

Ju-Hyun Bae[1], Yae-Won Sung[2], Ye-Eun Yuk[3], Yun-Hui Jang[4]
[1]Dept. of Computer Science and Engineering, Seoul National University of Science and Technology
[2]Dept. of Smart Information Technology, Bae-Wha Women's University
[3]Dept. of Division of Social Welfare, Kang-Nam University
[4]Dept. of Economics and Finance, Sang-Myung University

요 약

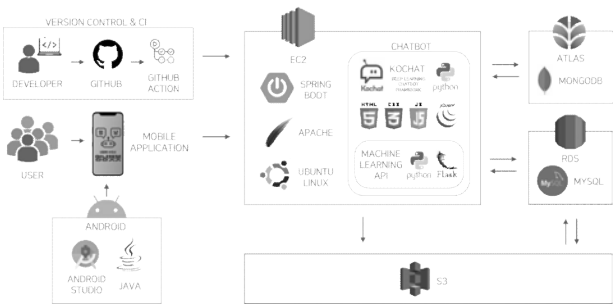
반려동물 시장 및 동물 의료분야의 성장, 동물병원 이용 과정 개선의 필요성으로 반려동물 질병의 시작부터 끝까지 전 과정을 함께하는 원스탑 모바일 애플리케이션을 개발하였다. 증상으로 예상 질병을 진단하는 머신러닝 모델과 자연어 문장을 인식하는 딥러닝 챗봇으로 사용자가 편리하게 반려동물 이상 증상에 대한 예상 질병을 챗봇으로 상담할 수 있도록 구현하였다. 챗봇 시스템을 기반으로 '예상 진단', '질병백과', '문진표', '동물병원' 기능을 추가하여 일관된 기능들로 유기적인 서비스를 구성하였다.

1. 서론

반려동물 시장은 팬데믹으로 인한 저성장 기조 속에서도 높은 성장세를 보이며 지속적으로 성장하고 있다. 특히, 코로나19의 확산으로 동물병원 내원에 어려움을 겪는 반려동물 양육자가 다수 발생하며 동물 의료분야가 급성장하였다. 우리나라의 경우 반려동물을 가족처럼 생각하는 문화적 현상이 강하게 나타나며, 펫펨족(Pet+Family), 딩펫족(Dink+Pet), 집사 등의 다양한 관련 신조어 생성으로 반려동물에 대한 사회의 높은 관심이 드러난다. 이에 '펫휴머니제이션', '펫테크', '첨단 동물의료' 키워드가 시장을 주도하는 새로운 반려동물 산업의 트렌드로 떠오르고 있다. 이와 함께, 동물병원 이용자의 만족도에 효율성이 가장 많은 영향을 미치는 것으로 나타났다. 정확, 간단, 편리, 신속할 수 있도록 [1] 동물병원 이용 과정의 개선이 필요하다. 따라서 사용자가 입력한 문장에서 증상을 추출하고, 예상 질병을 진단하는 챗봇 서비스를 기반으로 동물 질병을 치료하기 위한 모든 과정을 한 번에 처리할 수 있는 모바일 애플리케이션 개발하고자 한다.

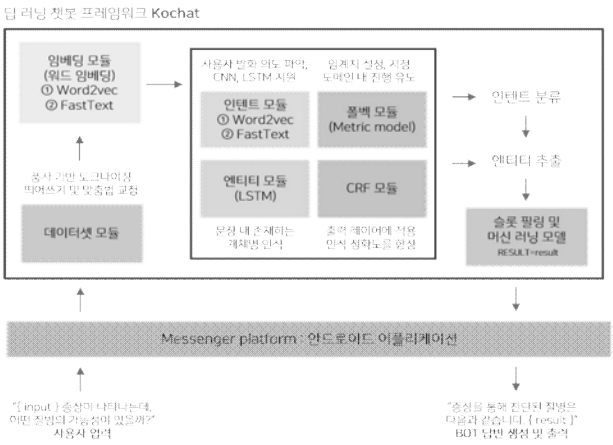
2. 본론

<그림 1> 시스템 아키텍처



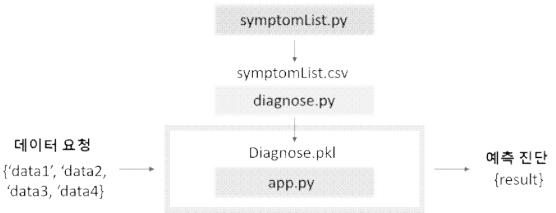
서비스의 시스템은 <그림 1>과 같이 구성한다. 사용자가 모바일 환경에서 안드로이드 애플리케이션으로 서비스를 이용할 수 있도록 구현한다. 챗봇 서비스는 딥러닝 챗봇 프레임워크, Kochat과 딥러닝 모델 CNN·LSTM을 활용하여 구현한다. Python 언어를 사용하여 딥러닝 챗봇과 머신러닝 API를 구현하며, HTML, CSS, JavaScript, jQuery를 활용하여 웹 페이지로 챗봇 기능을 구현하여 네이티브 애플리케이션에 웹뷰의 형태로 서비스한다.

<그림 2> 챗봇 아키텍처



사용자와의 상담으로 증상을 입력받고 예상 질병을 출력하는 챗봇은 <그림 2>와 같이 구성한다. 데이터셋 모듈에서 데이터를 전처리하고, 임베딩 모듈에서 FastText, Word2Vec을 통해 워드 임베딩한다. 인텐트 모듈에서 딥러닝 모델 CNN·LSTM의 지원으로 사용자 발화의 의도를 파악한다. 폴백 모듈에서 임계치를 정하고, 지정된 도메인 안에서 진행이 될 수 있도록 유도한다. CRF 모듈에서 CRF 출력 레이어에 CRF를 적용하여 인식 정확도를 향상한다. 엔티티 모듈에서 문장 내에 존재하는 개체명을 인식한다. 인텐트 모듈을 통해 슬롯을 고르고, 엔티티 모듈을 통해 해당 슬롯 필링을 한다.

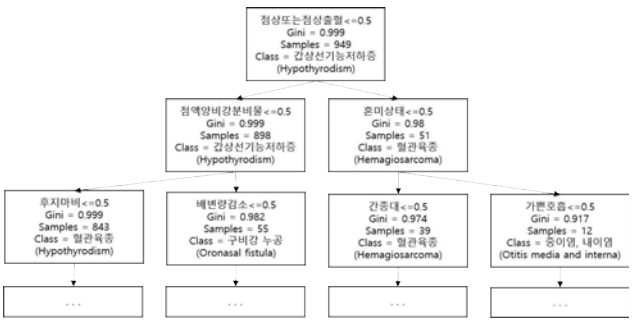
<그림 3> 머신러닝 API 아키텍처



예상 질병을 진단하는 머신러닝 API는 <그림 3>과 같이 구성한다. 실제 예시는 <그림 6>의 구현 내용으로 확인할 수 있다. 번호, 질병명, 축종, 정의, 원인, 발병기전, 발병및역학, 주요증상, 진단, 감별진단, 병리소견, 치료, 예방, 예후, 보호자에 대한 조언, 기타, 참고문헌으로 구성된 동물 질병 공공데이터 중 질병명, 축종, 주요증상 데이터를 판다스 라이브러리의 데이터프레임으로 불러온 후, 널 값을 제거하는 등 전처리를 수행한다. 증상을 단어 단위로 추출하여 'symptomList.py' 파일로 목록화한다. 주요증상 리스트를 원핫인코딩으로 벡터화한 'diagnose.py' 파일의 데이터로 머신러닝 모델이 학습되며, 생성된 모델을 'Diagnose.pkl' 파일로 저장한다. 'app.py' 파일로 API를 웹 프레임워크 Flask로 배포한다. 머신러닝 모델은 사이킷런 라이브러리의 DecisionTreeClassifier 또는 RandomForestClassifier를 사용한다.

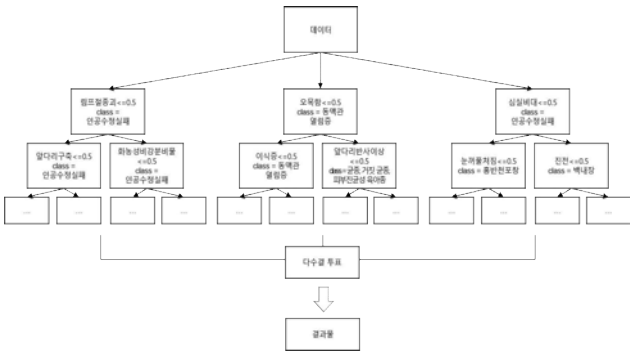
의사결정트리는 분류와 회귀가 모두 가능한 지도학습 머신러닝(machine-learning)의 일종으로 여러 클래스(class)로 분류된 기존 데이터를 토대로 새로 수집된 데이터가 어느 클래스에 속하는지 예측하는 판단규칙을 세우는 알고리즘이다. 의사결정트리의 장점은 시각적이고 명시적이며 결과의 해석이 직관적이고 쉽다는 데 있다. [2] 의사결정트리 모델을 진단 예측에 사용하기 위하여 동물 질병 공공데이터를 전처리하여 목록을 원핫인코딩으로 벡터화한다. 벡터화 후, 벡터의 값이 0인지 1인지에 따라 변수 영역을 두 개로 구분한다. 트리에 제한을 두는 파라미터의 값을 조정하여 과적합을 해결한다. Max depth를 규제하여 과적합을 피하고, 모든 value가 1인 루트 노트부터 시작하여 값이 0.5 이상인 경우(1)와 값이 0.5 이하인 경우(0)로 리프 노드를 분류한다.

<그림 4> 진단 예측을 위한
의사결정트리 모델 시각화의 일부 확대.



<그림 4>는 의사결정트리 모델의 일부를 max_depth 파라미터 값 2로 확대한 내용이다. 의사결정트리는 계단 모양의 결정 경계를 만들기에 훈련 세트의 회전에 민감하게 반응하여 훈련 데이터의 작은 변화에도 매우 민감하다는 문제가 발생한다. 이러한 단점을 보완하기 위해 랜덤포레스트를 사용할 수 있다. 랜덤포레스트는 여러 개의 결정트리를 생성하는 의사결정트리의 그룹이다. 의사결정트리는 의사 결정에 사용하는 규칙을 만드는 경향이 있는 반면 랜덤포레스트는 기능을 무작위로 선택하고 관찰하여 결정트리의 포레스트를 만들어 다수결 또는 평균을 계산하는 방식으로 결과를 도출한다. [3]

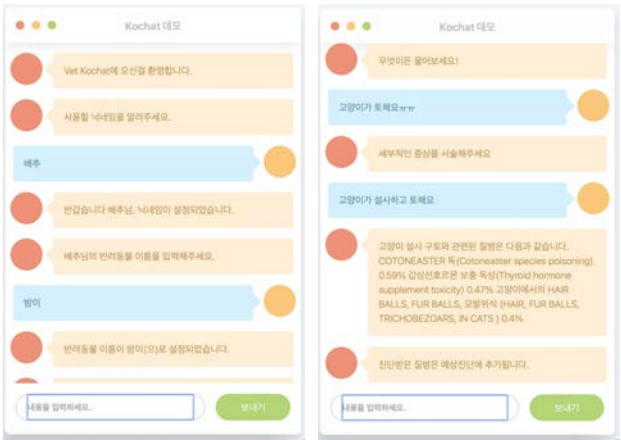
<그림 5> 진단 예측을 위한
랜덤포레스트 모델 시각화의 일부 확대.



따라서 본 연구에서는 의사결정트리의 과적합 문제를 보완하기 위한 해결책으로 랜덤포레스트 방식을 사용한다. 여러 개의 모델을 생성하고 다수결로 결과를 예측하는 랜덤포레스트 분류기를 사용한다.

3. 결과

<그림6> 반려동물 질병상담 챗봇 구현 화면



본 연구에서는 챗봇 시스템에 '예상 진단', '질병백과', '문진표', '동물병원' 기능을 추가하여 일관된 기능들로 유기적인 서비스를 구성하였다. '챗봇'으로 반려동물의 이상 증상 상담 후, AI로 진단받은 예상 질병을 '예상 진단'으로 관리, '질병백과'에서 검색으로 관련 정보를 확인할 수 있도록 하였다. 병원 내원이 필요한 경우, '문진표'에서 작성한 문진표를 활용하여 '동물병원'에서 지역별 검색을 통해 인근 동물병원을 찾아 연계병원에 예약으로 반려동물 질병의 시작부터 끝까지 전 과정을 함께하는 원스탑 모바일 애플리케이션을 구현하였다. 특히, 동물 질병 및 동물병원 공공데이터를 활용하여 구축한 데이터베이스로 '질병백과' 및 '동물병원' 기능을 통해 전문적인 지식을 사용자 편의에 맞춰 전달하고자 하였다. 챗봇의 경우, 증상으로 질병을 진단하는 머신러닝 API를 의사결정트리 모델과 랜덤포레스트 모델로 구현하여 Validation 데이터셋으로 성능 비교를 실시하였다. 그 결과 각 3.597%(개), 3.676%(고양이) 그리고 7.194%(개), 8.088%(고양이)의 정확도로 두 배 이상의 성능 차이를 확인되어 랜덤포레스트 모델을 채택하였다. 이에 더해, Grid search로 fine tuning을 진행하여 세부 파라미터를 조정함으로써 validation set에 대한 정확도를 13.971%(개), 13.669%(고양이)로 향상시켰다.

4. 결론

본 연구를 통한 사회·문화적 효과, 경제적 효과 및 기술적 효과가 기대되며, 본 연구는 높은 확장성과 상업성으로 다양하게 활용될 수 있다.

본 연구의 한계는 다음과 같다. 본 연구에서는 질병을 진단하는 머신러닝 모델의 학습을 위해 공공데이터를 활용하였다. 공공데이터는 한 가지의 클래스에 대해 한가지의 인스턴스를 가지고 있기에 cross validation을 사용할 수 없다는 한계를 가지고 있다. 따라서 네이버 지식인의 수의사 답변 크롤링을 통해 '증상으로 질병 진단'에 대한 개 140건, 고양이 137건의 데이터를 수집하여 validation 데이터로 사용했다. 하지만 같은 증상에 대해서도 여러 개의 질병 진단 가능성이 존재하여 동일한 feature에 다른 target을 갖는 데이터를 다수 수집하였고, 이에 의해 validation을 진행하였을 때 낮은 score가 도출되었다. 이는 데이터의 불충분에 의한 것으로 추후 지속적인 데이터 수집으로 개선할 예정이다.

참고문헌

[1] 조성민, 하규수. 반려동물병원의 이용만족에 미치는 영향에 관한 연구, 한국콘텐츠학회 논문지, 21(9), 258-269. 2021.

[2] 김경훈. 의사결정트리(Decision Tree)를 활용한 글로벌 부동산 가격 분석. 사회과학연구, 29(1), 107-122. 2022.

[3] 이주원, 홍성은, 방준일, 김화종. 기계학습 알고리즘을 사용한 미국 COVID-19 확진자 사망률 예측 연구. 동계종합학술발표회. 한국통신학회. 2021. 2 page.

※ 본 프로젝트는 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다.