TẠP CHÍ KHOA HỌC
TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
Tập 18, Số 6 (2021): 1146-1160
ISSN: 2734-9918
HO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
Vol. 18, No. 6 (2021): 1146-1160
Website: *http://journal.hcmue.edu.vn*

Bài báo nghiên cứu

NGHIÊN CỨU MÔ HÌNH HỆ THỐNG HỖ TRỢ TƯ VẤN CÔNG TÁC HỌC VỤ TRONG CƠ SỞ GIÁO DỤC ĐẠI HỌC

Phạm Nguyễn Huy Phương*, Vũ Thanh Nguyên,
Nguyễn Thị Diệu Hiền, Bùi Công Danh
Trường Đại học Công nghiệp Thực phẩm Thành phố Hồ Chí Minh, Việt Nam
**Tác giả liên hệ: Bùi Công Danh – Email: danhbc@hufi.edu.vn*
Ngày nhận bài: 15-3-2021; ngày nhận bài sửa: 17-5-2021; ngày duyệt đăng: 14-6-2021

TÓM TẮT

Chatbot là một hệ thống giao tiếp tương tác với con người bằng các phương pháp học máy, thực hiện cuộc trò chuyện thông qua một giao diện dưới dạng tin nhắn hoặc âm thanh. Trong thời kì chuyển đổi số ngày nay đã tạo điều kiện để chatbot tăng tốc nhanh chóng và tạo ra một hệ thống nhiều loại bot tương tự hệ sinh thái ứng dụng như trong việc chăm sóc khách hàng như cung cấp thông tin sản phẩm, đưa ra các thông tin gợi ý; quản lí hàng tồn, sắp xếp lịch, tra cứu dữ liệu y tế, chăm sóc sức khỏe. Trong bài báo này, chúng tôi nghiên cứu xây dựng một hệ thống chatbot có khả năng hỗ trợ tư vấn thông tin học vụ cho sinh viên bằng cách tiếp cận kết hợp các kĩ thuật gom cụm KNN, mạng nơron, mô hình túi từ và phương pháp thống kê TF-IDF. Bằng cách kết hợp các kĩ thuật máy học cũng như gom cụm, chúng tôi đã xây dựng được một mô hình tính toán cùng với một hệ thống tương tự chatbot để hiểu và trả lời những câu hỏi về thông tin học vụ.

Từ khóa: Chatbot; thuật toán KNN, ngôn ngữ tự nhiên; mạng nơron

1. Giới thiệu

1.1. Khái niệm

Chatbot là một hệ thống trao đổi thông tin giữa người và máy theo một quy chuẩn nhất định, thông tin trao đổi trong chatbot có thể bằng ngôn ngữ nói, ngôn ngữ viết hoặc kí hiệu.

Chatbot giúp cho người sử dụng tiết kiệm được thời gian, tiết kiệm chi phí trong việc ứng dụng vào các hệ thống chăm sóc khách hàng, hay nâng cao năng suất lao động hay thậm chí chăm sóc đời sống con người. Hệ thống chatbot được phân chia thành các loại chính như sau:

- Chatbot giữa con người với con người;
- Chatbot giữa máy tính với máy tính;
- Chatbot giữa con người và máy tính.

Cite this article as: Pham Nguyen Huy Phuong, Vu Thanh Nguyen, Nguyen Thi Dieu Hien, & Bui Cong Danh (2021). A model of a consulting assistance system for academic service in higher education. *Ho Chi Minh City University of Education Journal of Science, 18*(6), 1146-1160.

Như chúng ta đã biết, mạng xã hội lớn nhất thế giới Facebook đã giới thiệu về một nền tảng trao đổi tin nhắn vào năm 2016, với nhiều ưu điểm vượt trội như nền tảng thân thiện hơn, liên kết nhiều hệ thống và cho phép chúng ta có thể tạo riêng cho mình một hệ thống chatbot. Theo các nghiên cứu gần đây, tại Trung Quốc, WeChat là đơn vị tiên phong trong lĩnh vực này giới thiệu hệ thống chatbot Xiaoice – Chatbot khá hoàn thiện từ năm 2013 và đang ứng dụng hiệu quả trong nhiều lĩnh vực trong đời sống.

Có thể nói, Chatbot trong thời kì chuyển đổi số ngày nay không chỉ là dựa trên kịch bản đã sắp xếp trước mà còn được phát triển dựa trên nền tảng trí thông minh nhân tạo và máy học, chúng có khả năng tự học và tự phát triển cho phù hợp với thực tế. Nhiều nhà phân tích và dự báo cũng như các công ti công nghệ hàng đầu như Alphabet, Microsoft, IBM… đều đưa ra dự đoán Chatbot sẽ thống trị lĩnh vực dịch vụ khách hàng trong thời đại ngày nay đặc biệt là trong thời kì khủng hoảng bởi dịch bệnh Covid-19. Ví như như, theo (Pham, 2012) cho biết hệ thống chatbot đã hình thành và ra đời từ cách đây rất lâu. Cụ thể, vào năm 1950, ý tưởng của Turing là đưa ra một thiết bị thông minh sẽ thay thế con người thực hiện nhu cầu trao đổi thông tin, từ đó giúp hình thành nền tảng cho cuộc cách mạng về hệ thống chatbot. Tiếp theo đó, Eliza là chương trình chatbot đầu tiên được phát triển năm 1966 với mong muốn được tạo ra để trở thành nhà trị liệu tự động trả lời các câu hỏi đơn giản với các cấu trúc câu xác định. Hơn thế nữa, công trình của nhóm tác giả (Nguyen, & Truong, 2015) giới thiệu một phương pháp hỗ trợ công tác tư vấn tuyển sinh bằng cách sử dụng kĩ thuật học máy SVM kết hợp với hệ thống tin nhắn văn bản. Ứng dụng của nhóm tác giả đã mang lại những hiệu quả trong công tác tư vấn như tiết kiệm được nhân lực, thời gian và hệ thống trả lời tự động, tuy nhiên, công trình này còn nhiều hạn chế như chưa thu thập được các nguồn câu hỏi trên hệ thống khác. Gần đây, nhóm tác giả (Do, & Hoang, 2019) đã giới thiệu công trình xây dựng hệ thống chatbot hỗ trợ sinh viên ngành công nghệ thông tin trong việc tiếp cận xu hướng công nghệ trong lĩnh vực chuyên ngành, kĩ năng nghề nghiệp cũng như phương pháp học tập ở bậc đại học. Trong công trình này, nhóm tác giả sử dụng kết hợp nhiều phương pháp như k-láng giềng, mạng nơ ron, rừng ngẫu nhiên và máy véc tơ hỗ trợ để huấn luyện, phân lớp văn bản, tách từ, tìm câu trả lời phù hợp từ việc đặt câu hỏi bằng giọng nói của sinh viên thông qua hệ thống chatbot, kết quả thực nghiệm của công trình này cho kết quả độ chính xác khá cao.

1.2. Hệ thống chatbot

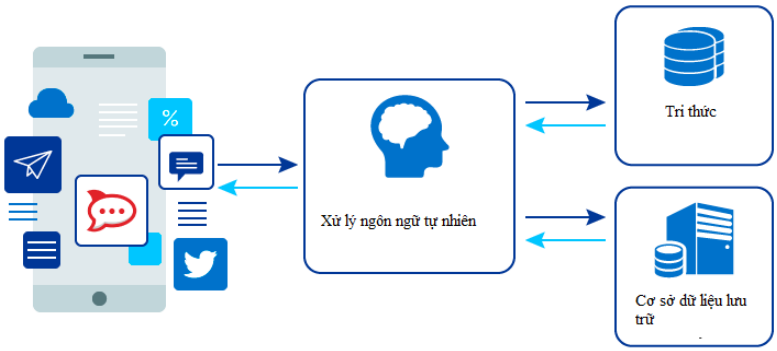
Kiến trúc cơ bản của một hệ thống chatbot bao gồm các thành phần cơ bản như sau:

- Cơ sở dữ liệu,
- Lớp ứng dụng,
- Quyền truy cập vào các API và giao diện đồ họa người dùng.

Cơ sở dữ liệu: Là nơi lưu trữ các loại thông tin, dữ liệu hoặc nội dung.

Tầng ứng dụng: Các giao thức trong tầng này được dùng để trao đổi thông tin giữa các chương trình chạy trên máy nguồn và máy đích. Tầng này có vai trò như là nơi xử lí các yêu cầu của các loại ứng dụng khác nhau.

Giao diện lập trình ứng dụng: là một giao diện mà một hệ thống máy tính cho phép các dịch vụ có thể được tạo ra từ các chương trình khác.



Hình 1. *Mô hình Chatbot* (Trương, & Ngo, 2014)

1.3. Phân loại

Chatbot kịch bản (Scripted chatbot): là chatbot có hành vi được xác định bởi các tiêu chuẩn, quy luật, trình tự. Tại mỗi bước trong cuộc trò chuyện, người dùng có thể thiết kế các trình tự nhất định theo nhu cầu sử dụng trong các ngữ cảnh khác nhau.

Chatbot thông minh (Intelligent Chatbot): là chatbot được xây dựng dựa trên nền tảng các kĩ thuật của máy học. Chúng cho phép người dùng cải thiện linh hoạt hơn về đầu vào có thể thu nhận đầu vào tự do dưới các hình thức sau: văn bản, giọng nói và cũng không giới hạn các dạng đầu vào khác nếu nó có ý nghĩa.

Một số ứng dụng triển khai chatbot

Subot: Subot hay gọi là trợ lí ảo trên Subiz, là một ứng dụng trên Subiz giúp tự động hóa kết nối Trả lời và Hỏi thông tin khách hàng. Từ đó, doanh nghiệp sẽ tăng tương tác với khách hàng 24/7 mà không bị phụ thuộc vào con người, chuyển đổi khách hàng tiềm năng bằng việc xin thông tin và xác định các yêu cầu cụ thể. Bạn (Agents) có thể xây dựng những kịch bản có sẵn cho Subot hoạt động như: Tự động trả lời khách hàng; Hiển thị đang nhắn tin; Hỏi thông tin liên hệ của khách hàng. Ngoài hỗ trợ Android Wear, còn có một ứng dụng Assistant cho iOS và dòng loa thông minh Google Home cũng sở hữu Google Assistant.

Simsimi: Simsimi là ứng dụng chat tự động rất thú vị trên di động. Được vào năm 2002, ISMaker – một phát triển phần mềm tại Hàn Quốc đã đưa ra ý tưởng đơn giản nhưng khá mạnh mẽ. Họ muốn tạo ra một ứng dụng đầu tiên có khả năng đưa tin theo dạng chatbot được đóng góp bởi cộng đồng. Vì vậy, Simsimi được ra đời với giao diện chú gà con màu vàng cực kì thân thiện. Người dùng có thể trò chuyện hoặc hỏi ứng dụng Simsimi bất cứ câu hỏi nào và nó sẽ thông qua cơ sở dữ liệu để đáp lại một cách ngẫu nhiên. Ý tưởng ban đầu

của nhà phát triển là mang đến những tiếng cười và niềm vui mỗi khi có người sử dụng Simsimi, điều đó được nêu trong bài viết trên blog của tác giả, và thời gian sau ứng dụng này đã phát triển nhanh chóng trở nên phổ biến rộng rãi với hàng triệu người dùng tại các quốc gia như Hàn Quốc, Thái Lan và Ấn Độ.

Miki: là một loại chatbot trên nền tảng Facebook. Chatbot này hỗ trợ nhiều tính năng chủ yếu về trên các loại lĩnh vực giải trí, tra cứu và học tập. Ứng dụng Miki có ưu điểm giúp người sử dụng không cần phải cài đặt thêm bất kì ứng dụng nào, chỉ cần bật Messenger và trao đổi thông tin với chatbot thì có thể sử dụng được trong việc tra từ điển Anh Việt, tra câu song ngữ Anh Việt cũng như dịch đoạn văn bản ngắn.

1.4. Đề xuất giải pháp chatbot trả lời tư vấn học vụ

Có thể nói, chatbot trở thành một hiện tượng mới trong việc đẩy mạnh chuyển đổi số trong tất cả các ngành nghề, doanh nghiệp và xã hội. Việc sử dụng chatbot trong các lĩnh vực như tiếp thị, quảng cáo của doanh nghiệp sẽ dễ dàng hơn và tiết kiệm chi phí. Bên cạnh đó, người dùng cũng cảm thấy hứng thú hơn và không còn cảm giác như đang phải bắt buộc tương tác với quảng cáo. Phù hợp với đa số người dùng và rất nhiều lĩnh vực, nếu như trước đây, chatbot chủ yếu trong các ngành nghề như bán lẻ, nhà hàng, bất động sản... thì giờ ngay cả các lĩnh vực cung cấp dịch vụ, hạ tầng, giáo dục, du lịch, chăm sóc sức khỏe cũng có thể sử dụng chatbot. Không chỉ dừng lại việc đáp ứng cho các doanh nghiệp lớn với lượng dữ liệu ngày càng tăng, chatbot còn phù hợp cho mọi mô hình doanh nghiệp bởi sự linh hoạt, dễ điều chỉnh theo nhu cầu. Sử dụng chatbot nhằm tiếp cận người dùng 24/7, đưa ra những phương án nhanh chóng, độ chính xác cao, giảm thiểu nhân công trong lĩnh vực tư vấn, hỗ trợ khách hàng. Từ đó, chatbot là công cụ hỗ trợ thúc đẩy doanh số, nâng cao hoạt động chăm sóc khách hàng, phát triển thương hiệu tốt nhất cho doanh nghiệp trong kỉ nguyên mới.

Trong bài bài báo này, chúng tôi trình bày hệ thống chatbot hỗ trợ trong công tác tư vấn học vụ tại các cơ sở giáo dục đại học bằng cách kết hợp xử lí ngôn ngữ tự nhiên và một số thuật toán như BOW, TF-IDF, ANN, KNN vào hệ thống để có thể phục vụ nhu cầu tìm kiếm thông tin một cách trực quan cho sinh viên, giảng viên và các bộ phận có liên quan. Kết quả thực nghiệm triển khai thực tế cho sinh viên Trường Đại học Công nghiệp Thực phẩm Thành phố Hồ Chí Minh và cho kết quả tốt hơn so với các phương pháp tiếp cận trước đây.

2. Vật liệu và phương pháp nghiên cứu

2.1. Mô hình túi từ

Trong thực tế ứng dụng, với một văn bản thì vector đặc trưng sẽ có dạng như thế nào, chúng ta có thể đưa các loại văn bản khác nhau về dạng vector nào cho phù hợp và theo nhiều phương pháp nghiên cứu trước đây sử dụng mô hình túi từ phù hợp nhất với vấn đề nêu trên. Mô hình túi từ là một biểu diễn đơn giản được sử dụng trong xử lí ngôn ngữ tự nhiên và tìm kiếm thông tin. Trong mô hình này, một văn bản được biểu diễn như là túi của các từ của nó, không quan tâm đến ngữ pháp và thậm chí cả thứ tự. Ví dụ, cho hai đoạn văn bản "ngành học A có điểm chuẩn cao hơn ngành học B" và "ngành học B có điểm chuẩn cao

hơn ngành học A" được biểu diễn giống nhau trong ngữ cảnh mô hình túi từ.

Theo nhóm tác giả (Do, & Tran, 2014) cho biết mô hình túi từ là một mô hình phổ biến cho biểu diễn dữ liệu dưới hình thức văn bản. Quá trình trích xuất đặc trưng của một văn bản bao gồm tách từ và đếm số lần xuất hiện của các từ trong văn bản. Như vậy, mô hình túi từ là mô hình biểu diễn văn bản như vector tần số xuất hiện của từ trong văn bản, được sử dụng phổ biến hiện nay trong vấn đề phân lớp văn bản thuộc lĩnh vực khoa học máy tính. Trong đó, từ điển được tạo thành từ tập tất cả các từ trong tập dữ liệu. Mỗi tài liệu (có thể là câu, đoạn hoặc văn bản) trong tập dữ liệu được biểu diễn dưới dạng vector đặc trưng, vector này có số chiều bằng với số từ có trong từ điển. Ví dụ, nếu tập dữ liệu có n từ thì vector của mỗi tài liệu trong tập dữ liệu sẽ có n chiều, mỗi từ khác nhau trong văn bản sẽ là một đặc trưng và tần số xuất hiện của nó trong văn bản là giá trị của đặc trưng tương ứng trí thành phần của vector là tần số xuất hiện của từ trong tài liệu.

Cũng theo nghiên cứu của nhóm tác giả (Do, & Pham, 2013) công bố mô hình túi từ, dữ liệu văn bản không có cấu trúcđược biểu diễn dưới dạng véc tơ tần số xuất hiện của từ trong văn bản, tập từ vựng trong tập dữ liệu có thể lên đến hàng chục ngàn, tập các dữ liệu văn bản được chuyển về dạng một bảng có số cột (chiều, từ vựng) rất lớn. Bên cạnh đó, công trình nghiên cứu của (Do, & Tran, 2014) cho biết nhược điểm của mô hình túi từ nằm ở chỗ không xác định đến sự đồng nghĩa của từ, điều này dẫn đến làm giảm hiệu quả dự đoán lớp dương hay lớp quan tâm của giải thuật k láng giềng trong phân lớp văn bản và có thể cho kết quả với độ chính xác không cao.

Nghiên cứu của hai tác giả (Do, & Pham, 2013) đã đề xuất phân loại văn bản bằng mô hình túi từ và mô hình máy học tự động dựa trên sự kết hợp giữa phương pháp biểu diễn văn bản bằng mô hình túi từ và các giải thuật xây dựng tập hợp các mô hình học tự động như Bayes thơ ngây ngẫu nhiên (random multinomial naive Bayes (rMNB)), cây xiên phân ngẫu nhiên đơn giản (random oblique decision stump (rODS). Các giải thuật boosting mới được đề xuất dựa trên mô hình cơ bản như cây ngẫu nhiên xiên phân đơn giản, Bayes thơ ngây ngẫu nhiên, cho phép phân lớp hiệu quả tập dữ liệu này. Kết quả thực nghiệm với tập dữ liệu thực cho thấy rằng phương pháp đề xuất phân lớp rất hiệu quả khi so sánh với các giải thuật hiện có, đạt được chính xác 94,8%".

Hơn thế nữa, theo nghiên cứu của tác giả (Do, & Tran, 2014) công bố phương pháp kết hợp ngữ nghĩa với mô hình túi từ để cải tiến giải thuật k láng giềng trong phân lớp văn bản ngắn. Trong bài báo này, nhóm tác giả đã giới thiệu tiếp cận tích hợp ngữ nghĩa với mô hình túi từ nhằm cải tiến hiệu quả dự đoán lớp dương của giải thuật k láng giềng trong phân lớp văn bản ngắn. Kết quả thực nghiệm với tập dữ liệu thực cho thấy rằng các phương pháp của nhóm tác giả đề xuất cải thiện dự đoán lớp dương hơn 8% trong khi giảm chưa đến 1% dự đoán lớp âm của giải thuật k-láng giềng trong phân lớp văn bản có độ dài ngắn. Ví dụ sau đây minh hoạ cách hoạt động của mô hình túi từ với tập dữ liệu văn bản như sau:

Bảng 1. *Ví dụ về tập dữ liệu văn bản*

STT	Nội dung
Tài liệu 1	Điểm chuẩn ngành dược
Tài liệu 2	Chỉ tiêu ngành dược
Tài liệu 3	Điểm trúng tuyển các ngành
Tài liệu 4	Xét tuyển theo học bạ

Từ tập dữ liệu trong Bảng 1, thu được từ điển gồm {điểm, chuẩn, ngành, dược, chỉ, tiêu, trúng, tuyển, các, xét, theo, học, bạ}. Từ điển trên gồm có 13 từ, vậy nên mỗi tài liệu sau khi véc tơ hóa sẽ có 13 chiều. Tần số xuất hiện của các từ được thể hiện như trong Bảng 2.

Bảng 2. *Biểu diễn tập dữ liệu bằng mô hình túi từ*

STT	điểm	chuẩn	ngành	dược	chỉ	tiêu	trúng	tuyển	các	xét	theo	học	bạ
Tài liệu 1	1	1	1	1	0	0	0	0	0	0	0	0	0
Tài liệu 2	0	0	1	1	1	1	0	0	0	0	0	0	0
Tài liệu 3	1	0	1	0	0	0	1	1	1	0	0	0	0
Tài liệu 4	0	0	0	0	0	0	0	1	0	1	1	1	1

Ta có các vector từ các tài liệu trong bảng 2 như sau:

- Vector của tài liệu 1: (1,1,1,1,0,0,0,0,0,0,0,0,0)
- Vector của tài liệu 2: (0,0,1,1,1,1,0,0,0,0,0,0,0)
- Vector của tài liệu 3: (1,0,1,0,0,0,1,1,1,0,0,0,0)
- Vector của tài liệu 4: (0,0,0,0,0,0,0,1,0,1,1,1,1)

2.2. Kĩ thuật TF-IDF

Khái niệm Term Frequency-Inverse Document Frequency, viết tắt là TF-IDF, thu được thông qua thống kê mức độ quan trọng của từ trong một văn bản, mà trong văn bản đang xét nằm trong một tập hợp nhiều văn bản đang xem xét. Giá trị TF-IDF tăng tương ứng với số lần một từ xuất hiện trong tài liệu, nhưng thường được bù đắp bằng tần số của từ trong kho văn bản, giúp điều chỉnh thực tế là một số từ xuất hiện thường xuyên hơn nói chung. Giá trị TF-IDF của từ t đối với văn bản d trong tập văn bản D là:

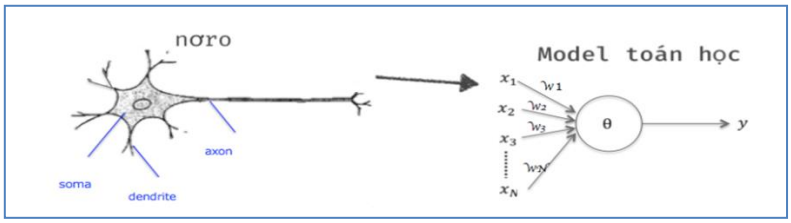
Tfidf(t, d, D) = tf(t,d) * idf(t, D)

với: - df(d, t): số lượng văn bản trong tập D có chứa từ t.

Các từ có giá trị TF-IDF cao là những từ xuất hiện nhiều lần trong văn này và xuất hiện ít trong văn bản khác, việc này giúp chúng ta lọc ra những từ phổ biến và giữ lại những từ có giá trị cao, nghĩa là từ khóa của văn bản đó.

2.3. Neural network

Neural là mô hình toán học mô phỏng nơron trong hệ thống thần kinh của con người. Mô hình biểu hiện cho một số chức năng của nơron thần kinh con người được mô tả như Hình 2



Hình 2. *Mô hình dây thần kinh nơron*

Tính chất truyền đi của thông tin trên neuron, khi neuron nhận tín hiệu đầu vào từ các dendrite, khi tín hiệu vượt qua một ngưỡng thì tín hiệu sẽ được truyền đi sang neuron khác theo sợi trục. Neural của model toán học ở đây cũng được mô phỏng tương tự như vậy. Công thức tính output Y sẽ như sau:

$$y=a\ (w^1x^1 + w^2x^2 + w^3x^3 - \theta) \quad (1)$$

với: y: tín hiệu output

x^1, x^2, x^3: tín hiệu input

w^1, w^2, w^3: weight

θ: ngưỡng threshold

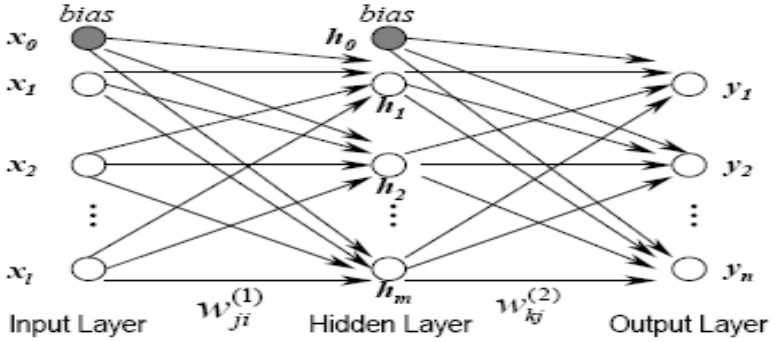
a: activation function

Thực tế threshold trong phạm vi toán học có thể mang dấu (+) và (-), dựa trên công thức (1) đưa vào công thức bias: bias = b = - θ. Suy ra được công thức sau:

$$y=a\ (w1x1 + w2x2 + w3x3+ b) \quad (2)$$

với: b: bias

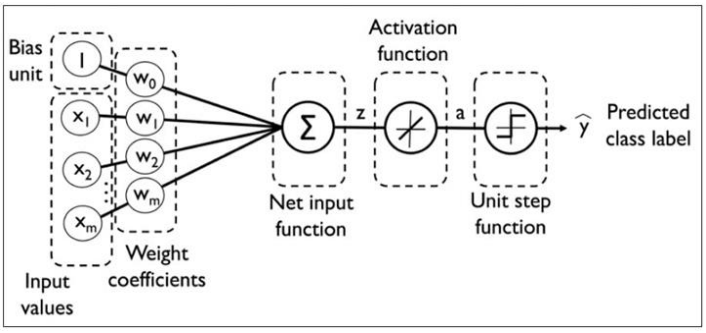
Một mạng nơ-ron là một tập hợp các nút nối với nhau, mô phỏng mạng nơ-ron thần kinh của não người. Mạng nơ-ron nhân tạo được thể hiện thông qua ba thành phần cơ bản: mô hình của nơ ron, cấu trúc và sự liên kết giữa các nơ ron. Trong nhiều trường hợp, mạng nơ-ron nhân tạo là một hệ thống thích ứng, tự thay đổi cấu trúc của mình dựa trên các thông tin bên ngoài hay bên trong chạy qua mạng trong quá trình học.



Hình 3. *Mạng nơron thần kinh*

Kiến trúc chung của một ANN gồm 3 thành phần đó là Input Layer, Hidden Layer và Output Layer

Một số cách thức thực hiện thuật toán học: Học tham số, học cấu trúc. Hai vấn đề này có thể được thực hiện đồng thời hoặc tách biệt. Nếu các mô hình, hàm chi phí và thuật toán học được lựa chọn một cách thích hợp thì mạng ANN sẽ cho kết quả có thể vô cùng mạnh mẽ và hiệu quả.



Hình 4. *Các thành phần của ANN*

Inputs (Đầu vào): Mỗi Input tương ứng với 1 đặc trưng của dữ liệu. Ví dụ như trong ứng dụng của ngân hàng xem xét có chấp nhận cho khách hàng vay tiền hay không thì mỗi input là một thuộc tính của khách hàng như thu nhập, nghề nghiệp, tuổi, số con...

Output (Đầu ra): Kết quả của một ANN là một giải pháp cho một vấn đề, ví dụ như với bài toán xem xét chấp nhận cho khách hàng vay tiền hay không thì output là yes/đồng ý hoặc no/không đồng ý.

Connection Weights (Trọng số liên kết): Đây là thành phần rất quan trọng của một ANN, nó thể hiện mức độ quan trọng, độ mạnh của dữ liệu đầu vào đối với quá trình xử lí thông tin chuyển đổi dữ liệu từ layer này sang layer khác. Quá trình học của ANN thực ra là quá trình điều chỉnh các trọng số Weight của các dữ liệu đầu vào để có được kết quả mong muốn.

Summation Function (Hàm tổng): Tính tổng trọng số của tất cả các input được đưa vào mỗi nơ-ron. Hàm tổng của một nơ-ron đối với n input được tính theo công thức sau:

$$Y = \sum_{i=1}^{n} X_i W_i$$

Transfer Function (Hàm chuyển đổi): Hàm tổng của một nơ-ron cho biết khả năng kích hoạt của nơ-ron đó còn gọi là kích hoạt bên trong. Các nơ-ron này có thể sinh ra một output hoặc không trong mạng ANN, nói cách khác rằng có thể output của một nơ-ron có thể được chuyển đến layer tiếp theo trong mạng nơ-ron hoặc không. Mối quan hệ giữa hàm tổng và kết quả output được thể hiện bằng hàm chuyển đổi.

2.4. Thuật toán KNN cho chatbot

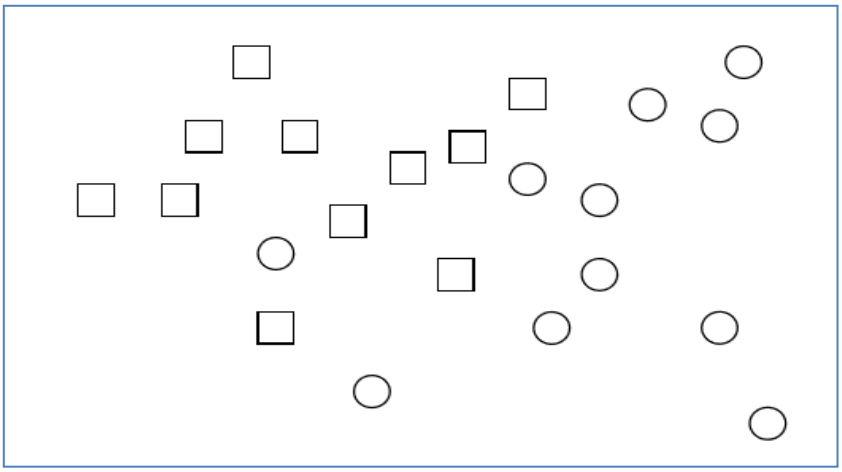
Một trong những phương pháp máy học thường được sử dụng để phân lớp và tìm kiếm văn bản là k láng giềng.

Giải thuật k-láng giềng (KNN – K Nearest Neighbors) được Fix và Hodges đề xuất từ những năm 1952. Đây là phương pháp rất đơn giản nhưng cũng cho hiệu quả cao trong khai mỏ dữ liệu. Giải thuật k láng giềng và phương pháp đánh giá hiệu quả phân lớp được mô tả chi tiết trong tài liệu. Phương pháp k-láng giềng (tên khác instance-based, lazy) rất đơn giản, dễ hiểu và thường cho kết quả tốt so với các phương pháp học khác. Giải thuật k láng giềng không có quá trình học, khi dự đoán lớp (nhãn) của phần tử dữ liệu mới đến, giải thuật đi tìm k láng giềng của nó từ tập dữ liệu học, sau đó thực hiện việc phân lớp phần tử mới đến. Quá trình phân lớp của k láng giềng mất rất nhiều thời gian. Giải thuật ứng dụng thành công trong hầu hết các lĩnh vực tìm kiếm thông tin, nhận dạng, phân tích dữ liệu, hồi quy (Do, 2017).

KNN là phương pháp để phân lớp các đối tượng dựa vào khoảng cách gần nhất giữa đối tượng cần sắp lớp và tất cả các đối tượng trong tập dữ liệu. Do quá trình tìm kiếm k phần tử lân cận cho mỗi phần tử mới, sau đó phân loại dựa trên luật bình chọn số đông (hồi quy dựa trên giá trị trung bình), độ phức tạp của quá trình phân loại khá lớn và kết quả phụ thuộc vào việc lựa chọn khoảng cách sử dụng.

Mục tiêu của các bộ máy – hệ thống tìm kiếm thông tin là trả về cho người dùng k tài liệu có độ tương đồng cao nhất so với nhu cầu thông tin của họ. Thực tế thì khi người dùng thực hiện truy vấn họ không biết được đâu là k tài liệu phù hợp với nhu cầu tìm kiếm của mình. Trong trường hợp này, hệ thống tìm kiếm sẽ cố gắng trả về k tài liệu có độ tương đồng cao nhất so với truy vấn từ người dùng. Trong bài báo này áp dụng phương pháp KNN để rút trích k tài liệu có độ tương đồng cao nhất với truy vấn của người dùng.

Ví dụ sau đây minh họa cách thức hoạt động của phương pháp KNN.



Hình 5. *Minh họa tập dữ liệu gồm 2 lớp*

Thuật toán KNN áp dụng vào bài báo được mô tả như sau:

- **Bước 1.** Để thực hiện bất kì thuật toán nào, chúng ta cần tập dữ liệu. Vì vậy, trong bước đầu tiên của KNN, chúng ta phải tải dữ liệu huấn luyện cũng như kiểm tra.
- **Bước 2.** Tiếp theo, chúng ta cần chọn giá trị của k tức là các điểm dữ liệu gần nhất. k có thể là bất kì số nguyên nào.

- **Bước 3.** Đối với mỗi điểm trong dữ liệu kiểm tra, hãy làm như sau:

- Tính toán khoảng cách giữa dữ liệu thử nghiệm và mỗi hàng dữ liệu huấn luyện với sự trợ giúp của bất kì phương pháp nào cụ thể là: Khoảng cách Euclidean, Manhattan hoặc Hamming. Phương pháp phổ biến nhất được sử dụng để tính khoảng cách là Euclidean.
- Sắp xếp khoảng cách trên theo thứ tự tăng dần.
- Chọn K hàng đầu tiên từ mảng đã sắp xếp.
- Chỉ định một lớp cho điểm kiểm tra dựa trên lớp thường xuyên nhất của các hàng này.

- **Bước 4.** Kết thúc

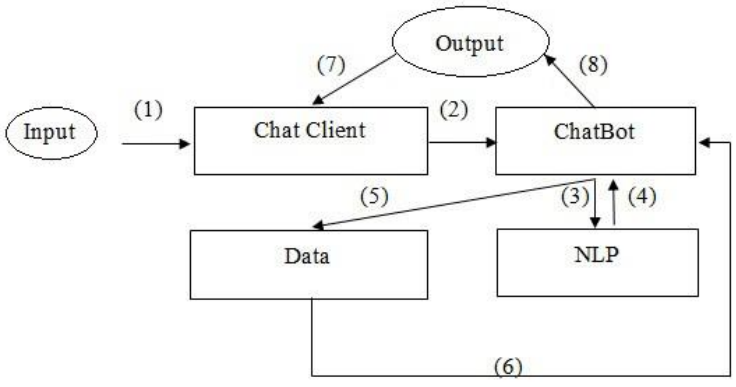
Việc tính toán khoảng cách giữa các đối tượng cần phân lớp với tất cả đối tượng trong tập dữ liệu huấn luyện thường được sử dụng với công thức tính khoảng cách Euclidean. Cho 2 điểm P1(x_1, y_1) và P2(x_2, y_2) thì khoảng cách Euclidean distance sẽ được tính theo công thức:

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

2.5. Đề xuất mô hình tư vấn học vụ

Hệ thống chatbot được xây dựng với mục đích ban đầu là đáp ứng nhu các yêu cầu cơ bản của một hệ thống tư vấn học vụ cho sinh viên tại Trường Đại học Công nghiệp Thực phẩm Thành phố Hồ Chí Minh hoặc các cơ sở giáo dục đại học. Dựa trên mô hình mạng neuron nhân tạo, mô hình túi từ và ứng dụng mô hình học máy này để xây dựng ứng dụng Chatbot hỏi-đáp. Kết quả thực nghiệm mô hình với tập dữ liệu thực cho thấy phương pháp của bài báo đề xuất là khá hiệu quả. Hệ thống chatbot thực nghiệm hoạt động có hiệu suất đúng như kì vọng.

Hệ thống chatbot với dữ liệu huấn luyện kịch bản hội thoại có sẵn, xây dựng giúp chatbot lấy được thông tin/câu hỏi từ phía người dùng. Chatbot xác định câu trả lời giúp người dùng có thể tiếp cận trực quan hơn với những câu hỏi mình muốn tìm kiếm. Cơ chế hoạt động của Chatbot như hình 6.



Hình 6. *Sơ đồ cơ chế hoạt động ChatBot*

Chú thích:

(1) Người dùng có câu hỏi dạng văn bản cần trả lời
(2) Người dùng nhập đoạn câu hỏi trên Chat Client
(3) Chatbot gửi đoạn câu hỏi về máy học
(4) Sử dụng NLP trích xuất các thông tin cần thiết của người dùng và gửi về cho Chatbot.
(5) Yêu cầu dữ liệu câu trả lời từ các thông tin cần thiết đã được xử lí.
(6) Dữ liệu câu trả lời được trả về cho Chatbot
(7) Chatbot gửi dữ liệu câu trả lời đến Chat Client
(8) Chat Client hiển thị câu trả lời cho người dùng.

3. Kết quả và thảo luận

3.1. Kết quả thực nghiệm

3.1.1. Dữ liệu thực nghiệm

Giới thiệu bộ dữ liệu: Bộ dữ liệu được thu thập và biên soạn tập dữ liệu từ website sinhvien.hufi.edu.vn của trường đại học công nghiệp thực phẩm bao gồm 286 câu hỏi và 293 câu trả lời liên quan đến các vấn đề như tư vấn học vụ, tham vấn học đường, kĩ năng mềm, chương trình đào tạo, sức khỏe, giáo dục…

3.1.2. Môi trường thực nghiệm

Để đánh giá hiệu quả của hệ thống chatbot đề xuất, nhóm tác giả cài đặt chương trình bằng ngôn ngữ lập trình Python. Để đảm bảo tính chính xác của chương trình, chương trình được chạy thực nghiệm trên các IDE Spyder3, Pycharm, Visual Studio Code có môi trường là anaconda3. Chương trình có sử dụng thư viện NLTK để thực hiện bước tách từ và biểu diễn các câu hỏi theo mô hình túi từ. Thư viện Scikit-learn được sử dụng để tạo bộ phận lớp KNN. Chương trình huấn luyện mạng nơron nhiều tầng. Thí nghiệm được chạy trên máy tính Acer Aspire 5 với CPU Intel core i5-7200 2.5Ghz 64bit, RAM 8GB, cài đặt hệ điều hành Windows 10.

- Cài đặt pycharm, spyder3, visual studio code
- Môi trường Anaconda 3

3.1.3. Quá trình thực nghiệm

Bước 1. Xử lí dữ liệu văn bản đầu vào

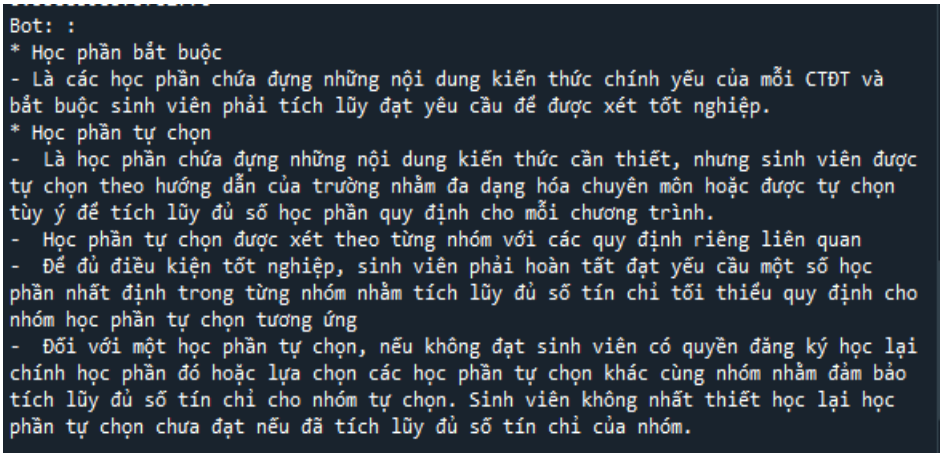
Input: Thế nào là các học phần bắt buộc, tự chọn

- Tách các từ trong câu thành từng từ đơn sử dụng thuật toán BoW
- Sử dụng thư viện nltk để giúp đỡ trong việc xử lí
- Bằng các thư viện như nltk.wordtokennize, nltk.stem
- Loại bỏ các kí tự không cần thiết
- Chuẩn hóa vector.

Bước 2. Xử lí phần thuật toán NeuralNet

- Sử dụng thư viện của NN để trainning và xác định dữ liệu đầu ra cho bài toán.

Bước 3. Xuất ra output



3.1.4. Kết quả so sánh số liệu giữa BoW và TF-IDF

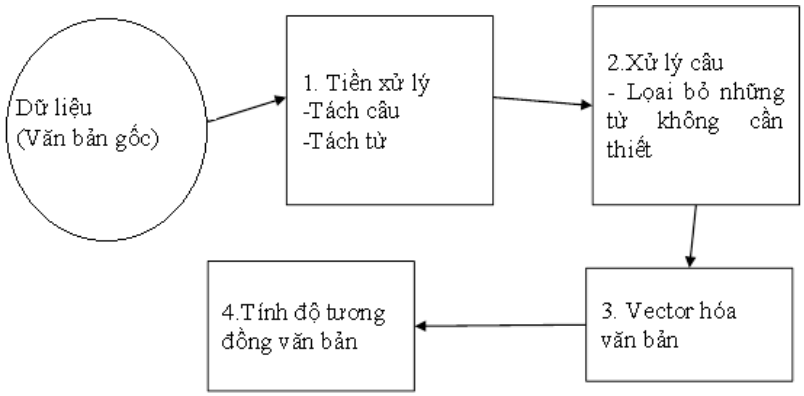
Để đánh giá mô hình BoW với KNN và TF-IDF với KNN, nhóm tác giả có sử dụng các chỉ số: k, Model, Distance Metric, Word Root, Accuracy để so sánh.

Bảng 3. *Bảng so sánh số liệu thuật toán bow và tfidf [6]*

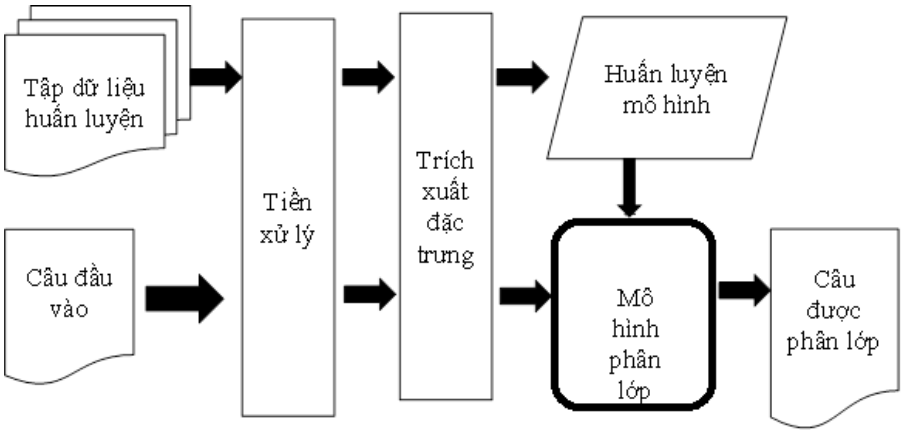
K	Model	Distance Metric	Word Root	Accuracy
1	BoW	Jaccard	Lem	64,50%
1	TF-IDF	Cosine	Lem	69,50%
1	BoW	Jaccard	Stem	70,50%
1	TF-IDF	Cosine	Stem	72,00%
5	BoW	Jaccard	Lem	71,00%
5	TF-IDF	Cosine	Lem	76,00%
5	BoW	Jaccard	Stem	70,50%
5	TF-IDF	Cosine	Stem	75,00%
10	BoW	Jaccard	Lem	68,00%
10	TF-IDF	Cosine	Lem	77,50%
10	BoW	Jaccard	Stem	69,50%
10	TF-IDF	Cosine	Stem	77,00%

Có thể dễ dàng nhận thấy được độ chính xác của giải thuật TF-IDF luôn cao hơn so với giải thuật BoW khi có cùng số k. Trong đó khi k bằng 10 thì giải thuật TF-IDF có độ chính xác cao nhất là 77,50% khi có word root là Lem.

3.1.5. Kết quả so sánh số liệu giữa ANN và KNN



Hình 7. *Sơ đồ mô hình hệ thống so sánh văn bản tiếng Việt*



Hình 8. *Sơ đồ mô hình huấn luyện phân lớp*

Để đánh giá mô hình cho bài toán, chúng tôi sử dụng các chỉ số:

Accuracy, k với k=10, hidden layer.

Bảng 4. *So sánh giữa ANN và KNN*

Các chỉ số	ANN	KNN
Accuracy	83,22%	76,58%

Bảng 4 cho thấy thuật toán ANN hoạt động hiệu quả tốt hơn, có độ chính xác cao hơn thuật toán KNN.

3.2. Thảo luận

Kết quả thực nghiệm cho chúng ta thấy độ chính xác của ANN hiệu quả hơn so với KNN nhưng vẫn còn có những tình huống mà hệ thống dùng ANN không thể trả lời câu hỏi có độ chính xác chưa cao thì hệ thống sẽ thông báo và lưu về hệ thống chờ quản trị viên cập nhật câu trả lời cho câu hỏi đó. Khi so sánh với thuật toán phân loại văn bản khác nhau như KNN, có thể thấy thuật toán ANN có độ chính xác cao hơn, kết quả thực nghiệm cũng cho kết quả tốt hơn. Kết quả vừa trình bày chưa phải là kết quả tối ưu, nhưng hi vọng rằng đây

sẽ là bước khởi đầu thuận lợi làm tiền đề nghiên cứu để thực hiện những chương trình trả lời tự động văn bản tiếng Việt tốt hơn nữa trong tương lai.

4. Kết luận

Trong bài báo này, nhóm tác giả trình bày nội dung tư vấn công tác học vụ tại cơ sở giáo dục đại học bằng phương pháp xây dựng Chatbot trên website trả lời tự động cho sinh viên các câu hỏi liên quan đến học vụ, các vấn đề về kĩ năng sống, môi trường, phương pháp học tập… Chatbot tư vấn học vụ được tạo dựa trên tiếp cận sử dụng máy học kết hợp với mô hình BOW và TF-IDF tạo một hệ thống hiệu quả giải quyết kịp thời nhu cầu của sinh viên và giảng viên. Hơn thế nữa, nhóm tác giả đã thu thập và biên soạn tập dữ liệu từ website của Trường Đại học Công nghiệp Thực phẩm Thành phố Hồ Chí Minh bao gồm 40 bộ dữ liệu hơn 286 câu hỏi và 293 câu trả lời khác nhau. Kết quả thực nghiệm cho thấy hệ thống đã có thể trả lời các câu hỏi mà người dùng hỏi với độ chính xác cao nhất là 83,45%.

- ***Tuyên bố về quyền lợi:*** *Các tác giả xác nhận hoàn toàn không có xung đột về quyền lợi.*
- ***Lời cảm ơn:*** *Nhóm tác giả cảm ơn Trường Đại học Công nghiệp Thực phẩm Thành phố Hồ Chí Minh đã hỗ trợ thực hiện công trình này.*

TÀI LIỆU THAM KHẢO

Do, T. N. (2017). *Giao trình Khai mo du lieu – minh hoa bang ngon ngu R* [*Data Mining - Illustrated in R language (Textbook)*]. Can Tho University Publishing House.

Do, T. N., & Pham, N. K. (2013). Phan loai van ban: Mo hinh tui tu va tap hop mo hinh may hoc tu dong [Text classification: a bag of word model and set of automatic machine learning models]. *Can Tho Univerisy Journal of Science, 28*(2), 9-15.

Do, T. N., & Tran, C. D. (2014). Ket hop ngu nghia voi mo hinh tui tu de cai tien thuat giai K lang gieng trong phan lop du lieu ngan [Combining semantic method with bag of word model to improve the K-neighbor algorithm in classifying short data]. *Can Tho Univerisy Journal of Science, 32*(1), 66-73.

Do, T. N., & Hoang, T. (2019). Chatbot cho sinh vien cong nghe thong tin [Chatbot for information technology students]. *Proceedings of conference on Fundamental and Applied IT research, Publishing House for Science and Technology*. doi: 10.15625/vap.2019.00012.

Nguyen, T. N., & Truong, Q. D. (2015). He thong ho tro tuyen sinh dai hoc [A consultancy support system for university entrance test]. *Can Tho Univerisy Journal of Science, CNTT (2015),* 152-159.

Pham, C. V. (2012). *Ung dung khai pha du lieu de tu van hoc tap tai truong cao dang kinh te – ki thuat quang nam* [*Apply data mining to support academic consulting at Quang Nam College Economics and Technology*]. Master's Thesis in Computer Science of The University of Danang, 1-25.

A MODEL OF A CONSULTING ASSISTANCE SYSTEM FOR ACADEMIC SERVICE IN HIGHER EDUCATION

Pham Nguyen Huy Phuong*, Vu Thanh Nguyen,
Nguyen Thi Dieu Hien, Bui Cong Danh
Ho Chi Minh City University of Food Industry, Vietnam
**Corresponding author: Bui Cong Danh – Email: danhbc@hufi.edu.vn*
Received: March 15, 2021; Revised: May 17, 2021; Accepted: June 14, 2021

ABSTRACT

A chatbot is a computer program or an artificial intelligence software that can interact with users in natural language, automatically simulate a conversation via an interface in the form of a message or sound. In the era of digital transformation, it has created conditions for chatbots to accelerate quickly and create a system of many types of bots similar to the ecosystem in customer care such as providing product information, offering suggestions, inventory management, scheduling, and medical data lookup and healthcare. In this article, we built a chatbot system capable of supporting academic counsulting for students by combining clustering method KNN, neural networks, bag-of-words model, and statistical measure TF-IDF. By combining machine learning and clustering techniques, we built a computational model with a chatbot system to understand and respond to questions related to academic affairs.

Keywords: chatbot; KNN; Natural Language; Neural Networks