ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**KHOA HỆ THỐNG THÔNG TIN**

------------------------



***Sinh viên thực hiện:***

1. Tạ Duy Thanh MSSV: 17521058

2. Võ Thị Duy Thảo MSSV: 17521071

3. Nguyễn Văn Trưa MSSV: 15520934

4. Nguyễn Thị Hồng Trúc MSSV: 17521168

***Giảng viên hướng dẫn:*** *Nguyễn Hồ Duy Trí*

---------------------------------------------------------------------------------

**ĐỀ TÀI: SENTIMENT CLASSIFICATION**

---------------------------------------------------------------------------------

BÁO CÁO ĐỒ ÁN MÔN HỌC KHAI THÁC DỮ LIỆU

LỚP: IS252.K13

Thành phố Hồ Chí Minh

LỜI CẢM ƠN

Đầu tiên, nhóm tác giả xin gửi lời cảm ơn chân thành đến tập thể quý thầy cô trường Đại Học Công Nghệ Thông Tin – Đại học quốc gia thành phố Hồ Chí Minh và quý thầy cô khoa hệ thống thông tin đã giúp cho nhóm tác giả chúng em có những kiến thức căn bản và chuyên sâu để thực hiện đề tài này. Đặc biệt, nhóm tác giả xin gửi lời cảm ơn chân thành và lòng biết ơn sâu sắc nhất đến thầy Nguyễn Hồ Duy Trí (Giảng viên thực hành môn khai thác dữ liệu). Các thầy cô đã trực tiếp hướng dẫn tận tình, sữa chữa và đóng góp nhiều ý kiến quý giá giúp nhóm tác giả hoàn thành tốt báo cáo môn học của mình. Trong thời gian một học kỳ thực hiện đề tài, nhóm tác giả đã sử dụng, vận dụng những kiến thức nền tảng đã tích lũy đồng thời kết hợp với việc học hỏi và nghiên cứu những gì đã thu thập được để hoàn thành một báo cáo đồ án tốt nhất. Tuy nhiên, trong quá trình thực hiện, nhóm tác giả không tránh khỏi những sai sót. Chính vì vậy, nhóm tác giả rất mong nhận được sự góp ý từ phía thầy cô nhằm hoàn thiện những kiến thức mà nhóm tác giả đã học tập và là hành trang để nhóm tác giả thực hiện tiếp các đề tài, dự án khác trong tương lai. Xin chân thành cảm ơn quý các thầy cô !

**Nhóm** **sinh viên thực hiện**

NHẬN XÉT CỦA GIẢNG VIÊN

Mục Lục

[LỜI CẢM ƠN 2](#_Toc27513754)

[NHẬN XÉT CỦA GIẢNG VIÊN 3](#_Toc27513755)

[**Phần 1: Lý do chọn đề tài** 5](#_Toc27513756)

[**Phần 2: Dữ liệu và miêu tả bài toán** 6](#_Toc27513757)

[*2.1. Dữ liệu (Dataset):* 6](#_Toc27513758)

[*2.1. Miêu tả bài toán:* 6](#_Toc27513759)

[**Phần 3: Kỹ thuật tiền xử lý dữ liệu** 6](#_Toc27513760)

[*3.1.Kỹ thuật tiền xử lý dữ liệu:* 6](#_Toc27513761)

[*3.2.Xử lý dữ liệu:* 7](#_Toc27513762)

[**Phần 4: Thuật toán khai thác dữ liệu** 8](#_Toc27513763)

[*4.1. Giải thuật K-nearest neighbor:* 8](#_Toc27513764)

[*4.2. Giải thuật Naïve Bayes:* 9](#_Toc27513765)

[*4.3. Deep learning:* 9](#_Toc27513766)

[*4.3.1 Word Vector* 9](#_Toc27513767)

[*4.3.2 Khảo sát tập dữ liệu train* 10](#_Toc27513768)

[*4.3.3 Mạng LSTM* 10](#_Toc27513769)

[**Phần 5: Kết quả đạt được** 11](#_Toc27513770)

[*5.1 :**Giải thuật K-nearest neighbor và Naïve Bayes* 11](#_Toc27513771)

[*5.2 :**Deep learning:* 11](#_Toc27513772)

[**Phần 6: Kết luận** 12](#_Toc27513773)

[**Phần 7: Bảng phân công công việc của các thành viên trong nhóm** 13](#_Toc27513774)

[*7.1. Bảng phân công :* 13](#_Toc27513775)

[*7.2. Bảng đánh giá :* 13](#_Toc27513776)

[**Phần 8: Tài liệu tham khảo** 14](#_Toc27513777)

# **Phần 1: Lý do chọn đề tài**

Thu thập thông tin phản hồi của khách hàng là một hình thức tiếp cận tuyệt vời và hữu ích giúp các doanh nghiệp hiểu được điểm mạnh, điểm yếu trong sản phẩm, dịch vụ của mình. Giúp doanh nghiệp hiểu rõ và nhanh chóng nắm bắt được tâm lý và nhu cầu của khách hàng để mang đến cho họ những sản phẩm, dịch vụ hoàn hảo nhất.

Hiện nay, với sự phát triển vượt bậc của khoa học và công nghệ, đặc biệt là sự bùng nổ của internet với các phương tiện truyền thông xã hội, những trang web, những app thương mại điện tử,… đã khiến cho mọi người không chỉ chia sẽ thông tin về sản phẩm mà còn thể hiện thái độ, quan điểm của bản thân về các sản phẩm, dịch vụ mà họ từng trải nghiệm.

Thông qua những dữ liệu được cung cấp qua internet bởi người dùng, mọi người có thể sử dụng nó để tìm kiếm, tham khảo trước khi đưa ra quyết định về việc lựa chọn sử dụng một sản phẩm hay dịch vụ nào đó. Đồng thời, Các nhà cung cấp dịch vụ cũng có thể sử dụng nguồn dữ liệu này để tự đánh giá về sản phẩm của mình, từ đó có thể đưa ra những cải tiến phù hợp hơn với người dùng. Mang lại lợi nhuận cao hơn, tránh các rủi ro đáng tiếc có thể xảy ra.

Nếu bạn đang sở hữu một trang thương mại điện tử hay một Blog,... thì một câu hỏi mà bạn chắc chắn đã từng nghĩ tới đó là thái độ, đánh giá và quan điểm của người dùng với sản phẩm bạn cung cấp để có thể đưa ra những chỉnh sửa, cải tiến cho hệ thống thích hợp hơn với người dùng. Tuy nhiên, để trả lời cho câu hỏi đó, cũng như giải quyết tất cả những vấn đề trên không phải là một chuyện đơn giản. Bạn có thể đi hỏi từng người sử dụng một, sử dụng phiếu đánh giá,... Nhưng thực tế cho thấy rằng những phương pháp này là không khả thi hoặc không hiệu quả cho lắm.

Để trả lời câu hỏi trên một cách hiệu quả và nhanh chóng thì bạn phải tận dụng phân tích được chính những thông tin mà người dùng để lại qua internet như các bình luận, đánh giá, bài chia sẻ,...

Sentiment Analysis là một chủ đề đầy tính thách thức trong Machine Learning

# **Phần 2: Dữ liệu và miêu tả bài toán**

# 2.1. Dữ liệu (Dataset):

Dữ liệu bài toán là những phản hồi của khách hàng tại những nhà hàng, quán ăn, tại các trang web thương mại điện tử chuyên review về ăn uống như foody, Lozi, bếp ăn uống,…

Link dataset : <https://streetcodevn.com/blog/dataset>

# 2.1. Miêu tả bài toán:

Giả sử bạn đang sở hữu một nhà hàng ăn uống tầm trung, và nhà hàng của bạn có trên những trang web đánh giá, review về ẩm thực, bạn có những mạng xã hội để PR cho nhãn hiệu của bạn như Facebook, instagram, twitter,…

Bạn muốn tìm hiểu xem khách hàng nói gì về nhà hàng của bạn để đánh giá, sửa chửa, cải thiện chất lượng của hiệu và bạn bắt đầu đi thu thập những comment liên quan trên mạng. Có những bình luận tích cực, và hiển nhiên cũng sẽ có bình luận tiêu cực, vấn đề ở đây là nếu số lượng comment quá nhiều thì bạn không thể kiểm soát hết được tất cả những comment đó.

Đây là lúc để bạn nghĩ đến việc xây dựng một model để làm việc đó thay bạn.

# **Phần 3: Kỹ thuật tiền xử lý dữ liệu**

# 3.1.Kỹ thuật tiền xử lý dữ liệu:

Đối với dữ liệu thu thập được bởi vì dữ liệu không sạch (có chứa lỗi, nhiễu, không đầy đủ, có mâu thuẫn…) thì các tri thức khám phá được sẽ bị ảnh hưởng và không đáng tin cậy, sẽ dẫn đến các quyết định không chính xác. Do đó, cần tiền xử lý dữ liệu; sửa chữa các dữ liệu nhiễu/lỗi; xác định hoặc loại bỏ các ngoại lai (outliers); giải quyết các mâu thuẫn dữ liệu. Các công việc cụ thể của tiền xử lý dữ liệu bao gồm những công việc như:

-Filtering Attributes: Chọn những thuộc tính phù hợp với mô hình

-Filtering Samples: Lọc các mẫu dữ liệu cho mô hình

-Clean Data: Làm sạch dữ liệu như xóa bỏ các dữ liệu bất thường

-Transformation: Chuyển đổi dữ liệu cho phù hợp với các mô hình như chuyển đổi dữ liệu từ numeric qua nominal hay ordinal

-Discretization: Nếu bạn có dữ liệu liên tục nhưng một số mô hình chỉ áp dụng cho các dữ liệu rời rạc (ví dụ như luật kết hợp) thì phải thực hiện việc rời rạc hóa dữ liệu

Các kỹ thuật tiền xử lý cơ bản được áp dụng vào đồ án:

-Chuẩn hóa về chữ thường

-Tách từ

-Loại bỏ dấu câu và ký tự đặc biệt

-Xử lý các trường hợp dùng láy âm tiết ( ngonnnn quáaaaa)

-Chuẩn hóa stopwords

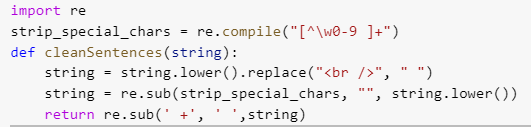
# 3.2.Xử lý dữ liệu:

Đối với dữ liệu thu thập được bởi vì dữ liệu không sạch (có chứa lỗi, nhiễu, không đầy đủ, có mâu thuẫn…) thì các tri thức khám phá được sẽ bị ảnh hưởng và không đáng tin cậy, sẽ dẫn đến các quyết định không chính xác. Do đó, cần tiền xử lý dữ liệu; sửa chữa các dữ liệu nhiễu/lỗi; xác định hoặc loại bỏ các ngoại lai (outliers); giải quyết các mâu thuẫn dữ liệu.

Các bước tiền xử lý dữ liệu gốc:

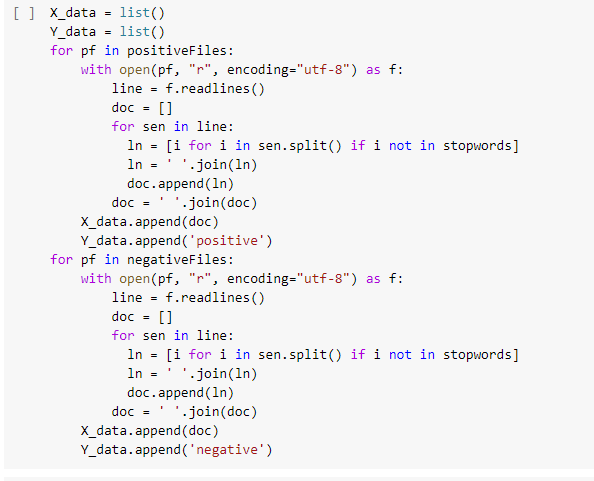
* Đọc dữ liệu từ các file text
* Tách tất cả các từ trong file đánh giá để lưu vào vector
* Loại bỏ những dữ liệu gây nhiễu/lỗi như:

Các ký tự đặc biệt | \ ? ' " ! ; : ( )….



Các chữ hay xuất hiện trong tiếng việt không mang nghĩa: a lô,

a ha, ai, ai nấy, ai đó , anh…



# **Phần 4: Thuật toán khai thác dữ liệu**

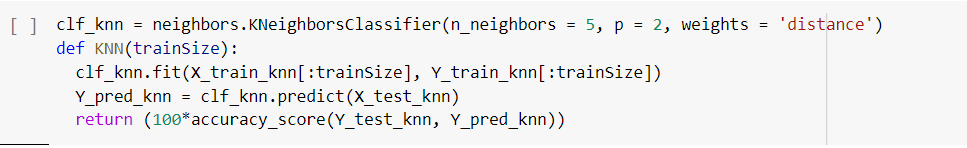
# 4.1. Giải thuật K-nearest neighbor:

Mỗi review được biểu diễn dưới dạng vector, trong đồ án này nhóm sử dụng TFIDF.

Sử dụng Norm 2 để tính khoảng cách giữa các vector.

Chọn K = 5.

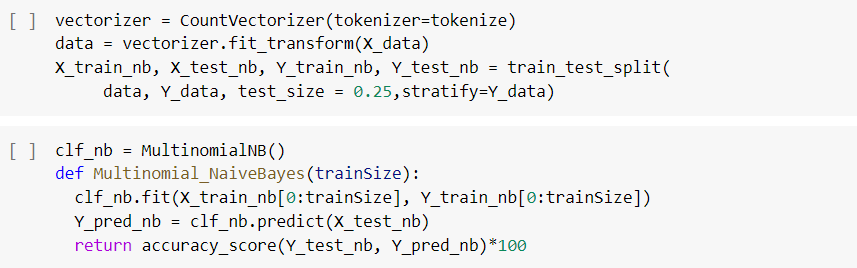
Đánh trọng số cho K điểm gần nhất, điểm càng gần điểm test thì được đánh số càng cao.



# 4.2. Giải thuật Naïve Bayes:

Nhóm sử dụng phân phối Multinomial Naïve Bayes

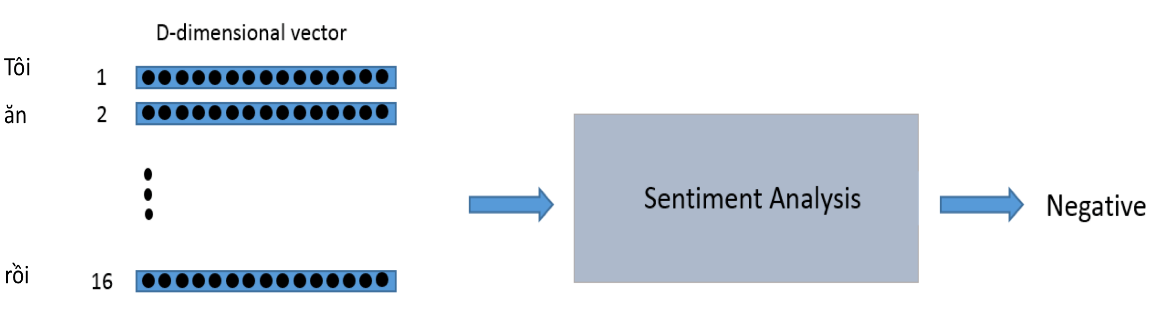
Các feature vectors được tính bằng Bags of words, mỗi văn bản được biểu diễn bởi một vector có độ dài d chính là số từ trong từ điển. Giá trị của thành phần thứ i trong mỗi vector chính là số lần từ thí I xuất hiện trong văn bản đó.



# 4.3. Deep learning:

4.3.1 Word Vector

Nếu như chúng ta giữ nguyên định dạng đầu vào là chuỗi ký tự thì rất khó để thực hiện các thao tác biến đổi như tích vô hướng ( dot product) hoặc các thuạt toán trên mạng neural network như backpropagation. Thay vì dữ liệu là một chuỗi, chúng ta cần chuyển đổi các từ trọng tập từ điển sang dạng vector số học.



Trong hình ở trên, ta có thể hình dung dữ liệu đầu vào của thuật toán phân tích cảm xúc là một ma trận 16 x D chiều. Trong đó 16 là số lượng từ trong câu và D là số chiều của không gian vector để biểu diễn từ. Để ánh xạ một từ sang một vector, chúng ta sử dụng ma trận word embedding (word2vec)

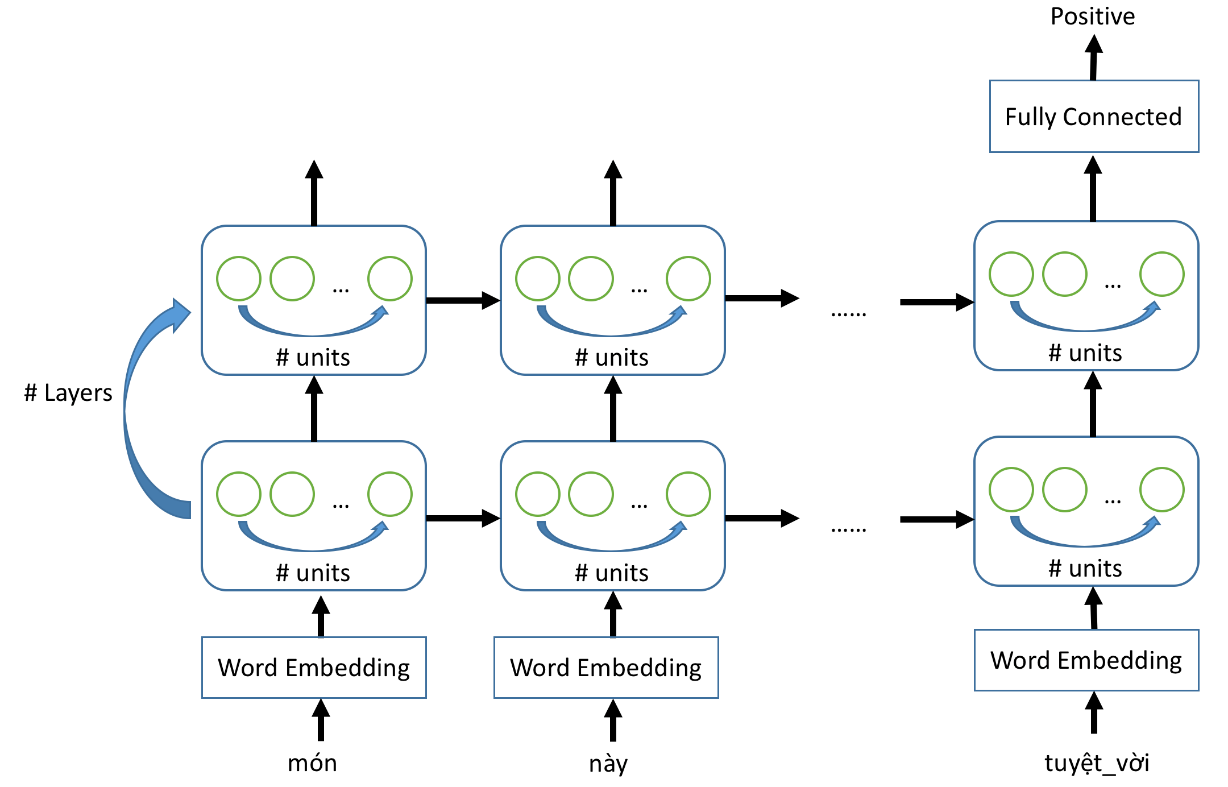
4.3.2 Khảo sát tập dữ liệu train

Do khối lượng dữ liệu lớn, nếu chúng ta chọn số lượng từ tối đa quá cao thì sẽ bị lãng phí khi biểu diễn ở những câu review quá ngắn. Ngược lại, nếu sử dụng số lượng từ tối đa quá ít thì sẽ bị bỏ lỡ những từ quan trọng giúp cho việc phân tích cảm xúc.

Sau khi thống kê độ dài của từng review thì chọn ra được độ dài tối đa để biểu diễn là 180 => mỗi input sẽ có dạng (180, 300) trong đó 180 là chiều dài của câu ( nếu câu nào không đủ thì các giá trị còn lại mặc định là 0), 300 là số chiều của vector biểu diễn từng từ.

4.3.3 Mạng LSTM

Để tăng tính phức tạp cho kiến trúc mạng chúng ta chồng các lớp LSTM lên nhau (Stac LSTM Layers). Trong trường hợp này chúng ta sử dụng 2 lớp LSTM. Việc chồng them các lớp LSTM sẽ giúp cho mô hình có khả năng nhớ nhiều thông tin hơn nhưng đồng thời cũng làm tang số lượng tham số khi huấn luyện. Điều này cũng có nghĩa là sẽ làm tang thời gian huấn luyện cũng như cần thêm nhiều dữ liệu hơn.

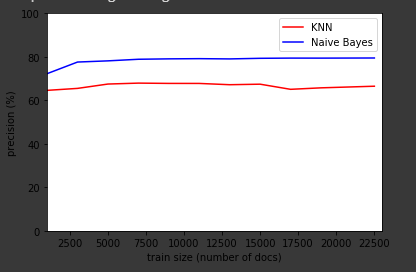


# **Phần 5: Kết quả đạt được**

# 5.1 :Giải thuật K-nearest neighbor và Naïve Bayes

Nhóm sử các kích thước tập train khác nhau để đánh giá 2 thuật toán lần lượt là: 1000, 3000, 5000, 7000, 9000, 11000, 15000, 19000, 22507, và kích thước tập test là : 7493

Kết quả thu được dưới dạng biểu đồ:

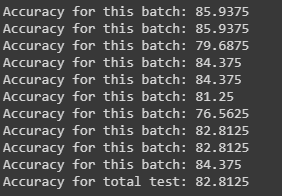


Ta nhận thấy với tập dữ liệu này, giải thuật naïve Bayes cho ra độ chính xác lớn hơn giải thuật K-nearest neighbor

# 5.2 :Deep learning:

Vì train mạng neuron khá tốn nhiều thời gian và tài nguyên. Cho nên nhóm chỉ train trên kích thước lớn nhất.

Và kết qua thu được trên tập test ngẫu nhiên với batch size = 64 có độ chính xác là:



…

# **Phần 6: Kết luận**

Nhìn tổng thể cho ta thấy khi ứng dụng Deep learning vào bài toán sẽ có kết quả cáo hơn một chút, nhưng đổi lại phải sử dụng tài nguyên và thời gian hơn vào việc huấn luyện mô hình

Tùy vào bài toán cụ thể ta sẽ chọn ra thuật toán để áp dụng

Hiện tại thì có các mô hình như BERT, ALBERT, XLNET… có độ chính xác rất tốt trên bài toán này, do thời gian khá ngắn cộng với lượng kiến thức còn khá hạn hẹp nên nhóm chưa tiếp cận được những phương pháp trên.

Hiện tại thì nhóm đang muốn phát triển theo hướng là thay vì mình dự đoán ra positive hay negative thì mình có thể đánh giá sao từ 1 đến 5 cho các review vì như thế sẽ có tính chi tiết cao hơn.

# **Phần 7: Bảng phân công công việc của các thành viên trong nhóm**

# 7.1. Bảng phân công :

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **MSSV** | **Công việc được giao** | **Mức độ hoàn thành (%)** |
| 1 | 17521058 | Viết báo cáo, phân tích thuật toán K-nearest-neighbor | 100% |
| 2 | 17521071 | Viết báo cáo, phân tích thuật toán Naïve Bayes | 100% |
| 3 | 15520934 | Viết Code, mạng neural Networks | 100% |
| 4 | 17521168 | Tiền xử lý dữ liệu | 100% |

# 7.2. Bảng đánh giá :

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **MSSV** | **Họ và Tên** | **Đánh giá chéo** |
| 1 | 17521058 | Tạ Duy Thanh | 10 |
| 2 | 17521071 | Võ Thị Duy Thảo | 10 |
| 3 | 15520934 | Nguyễn Văn Trưa | 10 |
| 4 | 17521168 | Nguyễn Thị Hồng Trúc | 10 |

# **Phần 8: Tài liệu tham khảo**

[1] <https://machinelearningcoban.com/2017/08/08/nbc/>

[2] <https://towardsdatascience.com/natural-language-processing-feature-engineering-using-tf-idf-e8b9d00e7e76>

[3] Assignment 3 – VietAI

[4] <https://streetcodevn.com/blog/dataset>