TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

Logo

Description automatically generated

**ĐỒ ÁN CUỐI KỲ MÔN XỬ LÝ**

**NGÔN NGỮ TỰ NHIÊN**

**XÂY DỰNG HỆ THỐNG PHÂN LOẠI Ý KIẾN NGƯỜI DÙNG DỰA TRÊN CÁC MÔ HÌNH MÁY HỌC**

*Người hướng dẫn*: **GV NGUYỄN TUẤN ĐĂNG**

*Người thực hiện:* **TRẦN MINH TRÍ** – **52000815**

Lớp : **200503401**

Khóa : **24**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

TỔNG LIÊN ĐOÀN LAO ĐỘNG VIỆT NAM

**TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

**KHOA CÔNG NGHỆ THÔNG TIN**

Logo

Description automatically generated

**ĐỒ ÁN CUỐI KỲ MÔN XỬ LÝ**

**NGÔN NGỮ TỰ NHIÊN**

**XÂY DỰNG HỆ THỐNG PHÂN LOẠI Ý KIẾN NGƯỜI DÙNG DỰA TRÊN CÁC MÔ HÌNH MÁY HỌC**

*Người hướng dẫn*: **GV NGUYỄN TUẤN ĐĂNG**

*Người thực hiện:* **TRẦN MINH TRÍ** – **52000815**

Lớp : **200503401**

Khóa : **24**

**THÀNH PHỐ HỒ CHÍ MINH, NĂM 2024**

LỜI CẢM ƠN

Em xin chân thành cảm ơn thầy Nguyễn Tuấn Đăng đã soạn những giáo trình, bài tập về Xử lý ngôn ngữ tự nhiên. Thông qua những kiến thức cùng với sự phổ cập của các tài liệu liên quan đến môn học đã giúp em dễ dàng hơn trong quá trình học. Trong quá trình thực hiện bài tập quá trình này sẽ không tránh khỏi sai sót, mong thầy góp ý và bỏ qua, em xin cảm ơn.

**BÁO CÁO ĐƯỢC HOÀN THÀNH**

**TẠI TRƯỜNG ĐẠI HỌC TÔN ĐỨC THẮNG**

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong Khóa luận/Đồ án tốt nghiệp còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

**Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung Báo cáo cuối kỳ của mình**. Trường Đại học Tôn Đức Thắng không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

*TP. Hồ Chí Minh, ngày 4 tháng 2 năm 2024*

*Tác giả*

*Trần Minh Trí*

**TÓM TẮT**

Ngày nay, với sự phát triển vượt bật của công nghệ kéo theo mức độ phổ biến của mạng xã hội cũng gia tăng nhanh chóng khiến cho con người càng ngày đắm chìm vào trong thế giới ảo này. Chúng ta sẵn sàng chia sẻ cảm xúc và ý kiến của mình trên các trang truyền thông, mạng xã hội, … để cho người khác biết mình đang nghĩ gì. Lợi dụng điều này, nhiều công ty đã và đang sử dụng dữ liệu từ các trang web khác nhau để tạo ra thông tin có ý nghĩa từ đó sử dụng vào mục đích kinh doanh của họ. Các dữ liệu văn bản khổng lồ có trên các trang như Google, Facebook, Amazon, IMDB,… là một nguồn tài nguyên vô cùng giá trị và việc phân tích chúng một cách thủ công là một lựa chọn vô cùng tồi tệ, vì nó quá chậm kèm theo nhiều rủi ro, sai sót,... Để giải quyết vấn đề đó, ta sẽ áp dụng các kỹ thuật lập trình để thu thập, trích xuất ý kiến của người dùng từ các phần bình luận, thảo luận, feedback,… và một trong số đó là kỹ thuật *Sentiment Analysis*. Đây là một nhánh con trong *Opinion Mining* mà trong đó việc phân tích sẽ tập trung vào giai đoạn khai thác, rút trích văn bản và ý kiến của người dùng dựa theo một chủ đề cụ thể. Trong đồ án này, ta sử dụng tập dữ liệu đánh giá về phim của IMDB và dùng các mô hình như Naïve Bayes, Support Vector Machine, và Logistic Regression để dự đoán liệu một phản hồi, bình luận của người dùng đánh giá phim là tiêu cực, tích cực hay trung lập. Ngoài ra, các phản hồi từ người dùng đa số sẽ là những câu văn mang thiên hướng sử dụng trong văn nói (gọi là *informal form*), do đó ta sẽ tiếp cận vấn đề bằng cách dùng *n-grams* và *count vectorizer*. Ta thực hiện *tokenization* để biến chuỗi đầu vào thành một véc-tơ từ, áp dụng *stemming* để xác định gốc của các từ, dùng *feature selection* để chọn các “đặc trưng” phù hợp và cuối cùng thực hiện phân loại (*classify*) dự đoán xem phản hồi người dùng mang khuynh hướng tích cực, tiêu cực hay trung lập.

**MỤC LỤC**

**DANH MỤC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ**

**CHƯƠNG 1 – TỔNG QUAN VỀ ĐỀ TÀI**

**1.1 Mục đích của đề tài**

Xem phim là một trong những nhu cầu giải trí của con người nhưng chỉ có một số ít bộ phim là thành công và được đánh giá cao. Có nhiều trang chuyên về đánh giá phim giúp người xem chọn lọc ra những bộ phim mà họ nên xem và ngược lại. Một trong số đó chính là IMDB, việc đánh giá mức độ thành công của bộ phim dựa vào số điểm trên thang 10 thông qua số sao mà người dùng đánh giá. Tuy nhiên, chưa có phương pháp nào thực hiện đánh giá bộ phim qua những phản hồi, bình luận. Vì thế, trong đề tài này, ta sẽ áp dụng kỹ thuật *sentiment analysis* để hiện thực hóa mục tiêu trên.

**1.2 Giới thiệu về Sentiment Analysis**

Sentiment Analysis (phân tích quan điểm) là quá trình giải nghĩa, phân loại cảm xúc trong văn bản sử dụng kỹ thuật phân tích văn bản. Phân tích quan điểm cho phép các doanh nghiệp xác định cảm nghĩ của người dùng đối với sản phẩm, thương hiệu hoặc dịch vụ thông qua các cuộc trò chuyện trực tuyến, phản hồi (feedback),… Các mô hình phân tích quan điểm không chỉ tập trung vào kết quả (tích cực, tiêu cực, trung lập) mà còn dựa trên cảm giác và cảm xúc (buồn, vui, tức giận,…). Phân tích quan điểm dần trở thành xu thế trong những năm gần đây và ngày càng có nhiều công ty lớn sẵn sàng đầu tư tài nguyên vào để dự đoán kết quả cho doanh nghiệp của họ.

Quá trình phân tích quan điểm gồm: *tokenization*, *word filtering*, *stemming* và *classification*. Trong quá trình tokenization, văn bản sẽ được phân đoạn thành các mẫu nhỏ thành các từ, các con số, hoặc các dấu câu. Kế tiếp dữ liệu sẽ được thực thi stemming để loại bỏ các tiền tố và phụ tố để chuyển một từ cụ thể thành từ gốc. Toàn bộ những bước trên còn được gọi là quá trình tiền xử lý dữ liệu, sau quá trình này ta sẽ phân tích tập dữ liệu bằng cách thực hiện phân loại sử dụng các phương pháp học máy như Naïve Baiyes, Support Vector Machine, và Logistic Regression. Cuối cùng ta sẽ chọn ra mô hình tốt nhất dựa trên độ chính xác của chúng, vì vậy ta cần phân tích và tìm ra những đặc trưng mà có ảnh hướng đến kết quả đánh giá phim để phân loại chúng là tích cực, tiêu cực hay trung lập.

**CHƯƠNG 2 – KHAI PHÁ DỮ LIỆU**

**2.1 Mô tả dữ liệu**

Bộ dữ liệu được thu thập và sử dụng trong đồ án này là bộ chứa 50000 reviews từ IMDB (Internet Movie Database), được chia đều thành hai tập training và testing (mỗi tập 25000 reviews). Cụ thể, trong mỗi phản hồi từ người xem sẽ chứa đựng ý kiến, nhận định và quan điểm của họ đối với một bộ phim cụ thể và với mỗi bộ phim ta chỉ có khoảng 30 reviews (lí do là vì các đánh giá trong cùng một bộ phim thường sẽ có độ tương quan nhất định). Ngoài ra, tập hợp các bộ phim của tập train và tập test là một tập không giao nhau (disjoint set), tức là trong quá trình huấn luyện mô hình thì có khả năng mô hình ghi nhớ một đoạn, một câu,… (gọi là specific terms) được liên kết với các nhãn nhất định sẽ không ảnh hưởng gì đến kết quả. Về việc gán nhãn ở bước cuối, ta sẽ dựa vào số điểm (score) sau khi dự đoán review để gán nhãn như sau:

score ≤ 4: negative

score ≥ 7: positive

4 < score < 7: neutral

**2.2 Tiền xử lý dữ liệu**

Để mô hình đạt được hiệu suất tốt, ta cần thực hiện một số phương pháp để sàn lọc dữ liệu nhằm loại bỏ những dữ liệu bị nhiễu (noise), tức những dữ liệu không cần thiết để giúp quá trình phân loại chính xác hơn. Cụ thể trong báo cáo này, đối với dữ liệu ta đang xử lý thì nó bao gồm các bước như sau: loại bỏ các ký tự HTML, lemmatization, loại bỏ stop words và các ký tự đặc biệt, và tokenize.

***2.2.1 Loại bỏ các thẻ HTML***

Trong hầu hết các trường hợp khi dữ liệu được thu thập (crawling) từ các trang web sẽ không tránh khỏi việc sẽ có những thẻ HTML dư thừa trong văn bản. Điều này có thể ảnh hướng đến kết quả, độ chính xác của mô hình. Bởi vì, các thẻ HTML như (<div>, <p>, <a>, …) vốn dĩ không thuộc phạm trù về mặt ngôn ngữ tự nhiên (ta không sử dụng chúng trong văn nói, văn viết). Do đó, loại bỏ các ký tự này sẽ giúp văn bản “clean” hơn, và cũng đảm bảo cho các bước xử lý, phân tích phía sau trở nên “trơn tru” hơn vì chúng hoàn tập tập trung xử lý trên những văn bản có mang tính ngữ cảnh thay vì những văn bản có đánh dấu in đậm, in nghiêng, được format,…

Để thực hiện loại bỏ các thẻ HTML, ta sẽ dùng regex (regular expressions) để tạo một pattern khớp với các từ nằm trong hai dấu “<” và “>” như đoạn code sau đây:



trong đó:

* “<.\*?>”: Tìm tất cả ký tự mà bắt đầu bằng dấu “<”, kết thúc bằng dấu “>” và “.\*?” nghĩa là tồn tại 0 hoặc vô số bất kỳ ký tự nào (trừ ký tự xuống dòng)
* Sau đó thay thế toàn bộ những ký tự khớp pattern trên bằng ký tự rỗng (empty string).

***2.2.2 Lemmatization***

Lemmatization (bổ đề hóa) là một quá trình rút gọn các từ về dạng gốc của chúng (được gọi là bổ đề). Ví dụ:

* Bổ đề của từ “running” là “run”.
* Bổ đề của từ “better” là “good”.

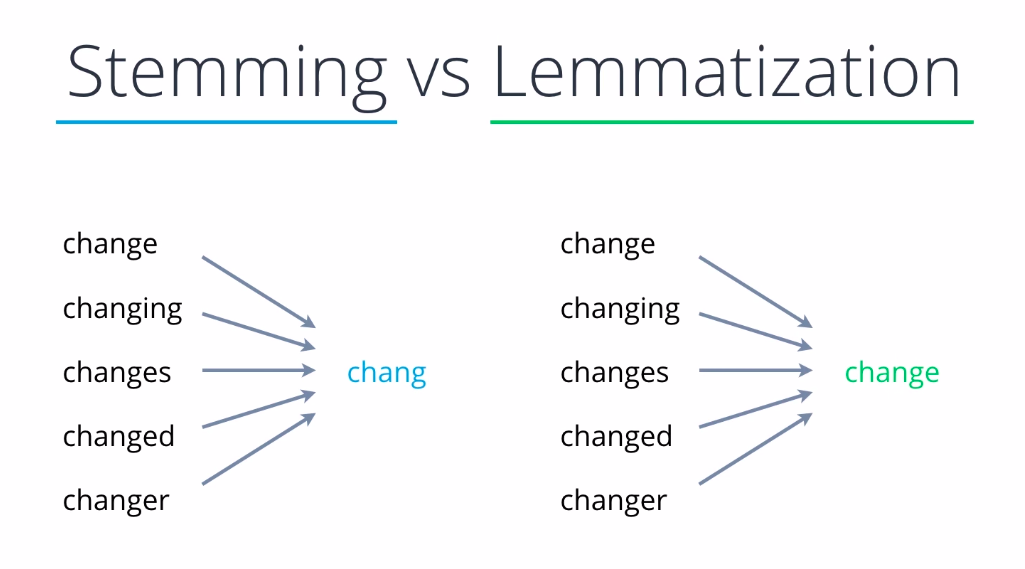
Mặc dù có cùng mục đích là để rút gọn các từ về dạng cơ bản của chúng. Song, đối với bài toán phân tích quan điểm, ta thường “ưa chuộng” lemmatization hơn là stemming là bởi vì ta cần quan tâm đến ngữ cảnh, ta muốn mô hình “hiểu” được ngữ cảnh của câu văn trong quá trình huấn luyện. Mà cơ chế của stemming là rút gọn bằng cách sử dụng các rules (thuật toán Porter năm 1980) như sau:

\begin{example}
\begin{tabular}[t]{lll@{\hspace{1in}}lll}
\multicolumn{3}{l}{\te...
... & $\rightarrow$\ & & cats & $\rightarrow$\ & cat \\
\end{tabular}\end{example}

\begin{figure}\raggedright
\begin{description}
\item[\emph{Sample text:}]
Such a...
...at is mor biolog transp and access to
interpret
\end{description}
\end{figure}

Hình: Ví dụ về stemming.

Ta thấy, stemming chỉ tập trung vào việc lược bỏ các prefix và suffix, điều này dẫn tới việc mất tính ổn định về mặt ngữ pháp (Part-of-Speech). Lấy ví dụ, từ “running” đóng hai vai trò vừa là động từ và danh từ, nhưng cơ chế của stemming đều sẽ đưa cả hai về dạng gốc là “run”, còn đối với lemmatization động từ “running” sẽ được đưa về “run” trong khi danh từ “running” thì vẫn được giữ nguyên. Tóm lại, nếu cân nhắc về khía cạnh ngữ cảnh của từng từ thì sử dụng lemmatization sẽ tốt hơn



Hình: Sự khác nhau giữa stemming và lemmatization.

Trong đồ án này, ta sử dụng hàm WordNetLemmatizer từ thư viện nltk để thực hiện lemmatize các từ:



***2.2.3 Loại bỏ stop words và các ký tự đặc biệt***

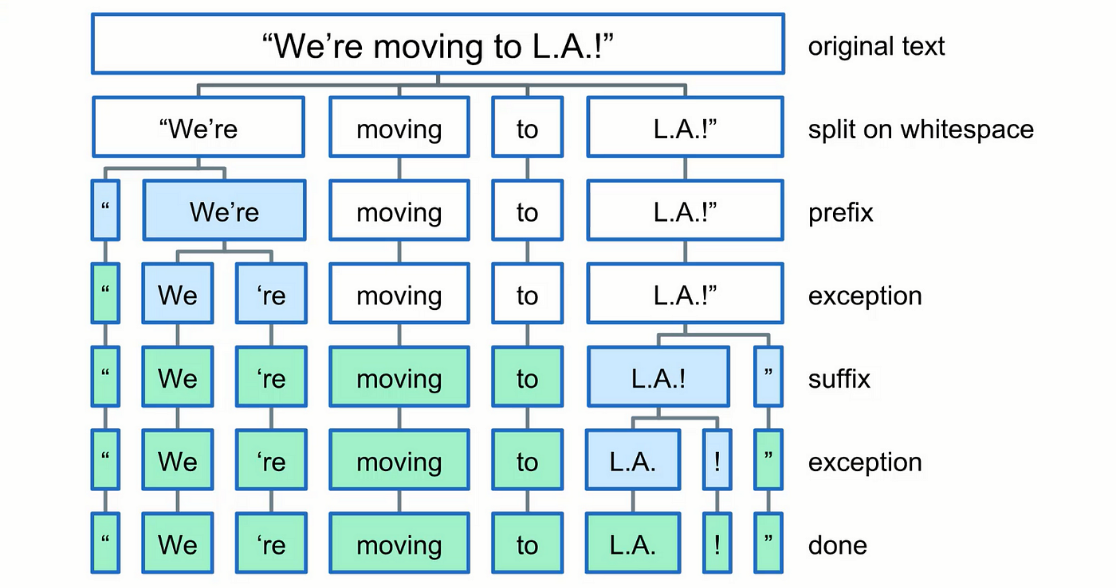
Trong xử lý ngôn ngữ tự nhiên, stop words là những từ không mang nhiều ý nghĩa, và việc có hay không có chúng cũng chẳng ảnh hưởng gì đến ý nghĩa của câu. Dĩ nhiên, đối với từng loại ngôn ngữ khác nhau thì stop words cũng sẽ khác nhau nhưng về mặt bản chất thì chúng thường là những mạo từ, dấu câu (ví dụ: trong Tiếng Anh thì các mạo từ là ‘a’, ‘an’, ‘the’, ‘this’, ‘that’, ‘these’, ‘those’,… là các stop words). Việc loại bỏ các stop words sẽ giúp mô hình khi huấn luyện sẽ chỉ quan tâm đến những từ có ý nghĩa, không bị ảnh hưởng bởi các dữ liệu nhiễu giúp tăng hiệu suất, cũng như độ chính xác của mô hình. Lấy ví dụ, ta có câu “The quick, brown fox jumps over the lazy dog!” có thể được viết rút gọn lại thành “quick brown fox jumps lazy dog”. Tức ta có thể loại bỏ đi những mạo từ, dấu câu (“the”, “over”, “!”) mà vẫn giữ nguyên được ý nghĩa của câu.

Trong báo cáo này, ta sẽ dùng bộ stop words trong tiếng Anh được cung cấp bởi thư viện nltk. Sau đây là hàm dùng để loại bỏ các stop words và các ký tự đặc biệt:



***2.2.4 Text tokenization***

Text tokenization (mã hóa văn bản) là quá trình chia nhỏ cụm từ, câu văn, hoặc đoạn văn,… thành các mẫu nhỏ hơn được gọi là token. Mỗi token có thể là một ký tự, một dấu câu, một từ, một câu văn,… tùy thuộc vào từng yêu cầu của bài toán cụ thể. Sau đây là ví dụ minh họa về quá trình mã hóa:



Hình: Quá trình tokenize ([media.com](https://www.google.com/url?sa=i&url=https%3A%2F%2Fmedium.com%2F%40utkarsh.kant%2Ftokenization-a-complete-guide-3f2dd56c0682&psig=AOvVaw1RMpqXSj2JBf2FFp-WF4P-&ust=1706060883226000&source=images&cd=vfe&opi=89978449&ved=0CBQQjhxqFwoTCPDaypqy8oMDFQAAAAAdAAAAABAQ))

Mục đích của việc mã hóa là để giúp mô hình hiểu cấu trúc của ngôn ngữ cấu thành văn bản. Việc chia nhỏ câu thành từng đơn vị để dễ dàng phân tích ngữ nghĩa, cấu trúc cú pháp và các mối quan hệ ngữ nghĩa từ đó giúp tăng độ hiệu quả trong việc indexing và searching (đối với bài toán information retrieval) và tiết kiệm nguồn tài nguyên của máy tính.

Ở đây ta sử dụng thư viện NLTK vì nó hỗ trợ rất nhiều ngôn ngữ như Tiếng Anh, Pháp, Tây Ban Nha,… Ngoài ra, trong NLTK, word tokenization là một hàm wrapper sử dụng treebank tokenization và nó phân tách cả các dấu câu (tức mỗi dấu câu cũng là một token, ngoại trừ dấu chấm).

Sau đây là quá trình thực hiện tiền xử lý dữ liệu đối với bài toán phân tích quan điểm với bộ dữ liệu IMDB, trong đó bước cuối ta sẽ thực hiện tokenize thông qua hàm có sẵn trong thư viện nltk:



**2.3 Trích xuất đặc trưng**

Cũng như bao lĩnh vực khác như thị giác máy tính, học sâu,… trích xuất đặc trưng là một bước vô cùng thiết yếu. Đối với các bài toán trong xử lý ngôn ngữ tự nhiên, đặc biệt là phân tích quan điểm người dùng thì mục tiêu của quá trình này là chuyển dữ liệu đầu vào về dạng biễu diễn số để đưa vào các mô hình huấn luyện. Một trong những hướng tiếp cận cho vấn đề này là sử dụng mô hình Bag-of-Words và kỹ thuật Count Vectorizer để triển khai mô hình này, và một số kỹ thuật cao hơn như TF-IDF(Term frequency-Inverse Document Frequency), các phương pháp word embeddings như Word2Vec, Glove,…

***2.3.1 Mô hình túi từ sử dụng Count Vectorizer***

Mô hình túi từ (Bag of Words) là một mô hình thể hiện một văn bản hoặc tài liệu dưới dạng một tập hợp các từ không có thứ tự, không quan tâm đến ngữ pháp mà chỉ quan tâm đến tần suất của các từ.

Cụ thể, ta sẽ tạo một tập hợp chứa toàn bộ các từ (mỗi từ là duy nhất) có trong tập ngữ liệu (corpus). Mỗi văn bản, tài liệu sẽ được biểu diễn dưới dạng véc-tơ, trong đó mỗi phần tử tương tứng với số lượng xuất hiện của phần tử đó trong văn bản, tài liệu đang xét. Lấy ví dụ, ta có hai câu như sau:

1. “I love NLP”
2. “NLP is fascinating”

thì túi từ của chúng sẽ có dạng như sau: [1, 1, 1, 1, 0, 0, 0, …], trong đó mỗi vị trí sẽ tương ứng với một từ độc lập trong tập từ vựng. Một số thư viện trong python có hỗ trợ ta thực hiện điều này, sau đây là một ví dụ sử dụng sklearn:



***2.3.2 TF-IDF***

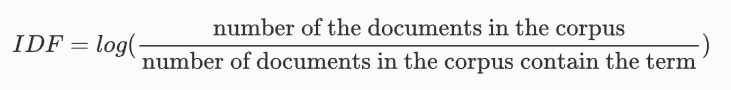
Cũng giống như BoW, TF-IDF là một kỹ thuật thống kê giúp phản ánh tầm quan trọng của một từ trong tập tài liệu (document) so với tập hợp các tập tài liệu (corpus). Nó là sự kết hợp giữa hai phần, term frequency và inverse document frequecy.

Tần suất thuật ngữ (term frequency) thực chất là tính số lần xuất hiện của một thuật ngữ trong một tài liệu chia cho tổng số thuật ngữ trong tài liệu đó.

A black text on a white background

Description automatically generated

Inverse document frequency dùng để đánh giá mức độ quan trọng của một từ trong văn bản, tuy nhiên trong văn bản thông thường thì những từ không quan trọng lại có xu hướng xuất hiện với tần suất cao (ví dụ: and, or, in, at, this, that,…). Do đó, ta cần giảm mức độ quan trọng của những từ đó bằng cách sử dụng IDF.



Những từ có TF-IDF cao là những từ xuất hiện trong văn bản này nhiều và xuất hiện ít trong văn bản khác. Việc này giúp lọc ra những từ ngữ phổ biến và giữ lại những từ có giá trị cao trong văn bản

A black text on a white background

Description automatically generated

Tuy nhiên, cũng giống như BoW thì TF-IDF chỉ quan tâm đến tần suất xuất hiện của các từ trong văn bản để đánh giá xem từ nào là thực sự quan trọng, từ nào không nhưng đối với bài toán phân tích quan điểm thì nó lại thiếu một yếu tố quan trọng đó là mối quan hệ ngữ nghĩa và điều này sẽ được khắc phục phần nào thông qua kỹ thuật Word2Vec.

***2.3.3 Word2Vec***

**CHƯƠNG 3 – MÔ HÌNH PHÂN LOẠI**

*Sơ lược: Trong báo cáo này, ta sử dụng ba mô hình gồm: Naïve Bayes, Logistic Regression và Support Vector Machine. Ta sẽ xây dựng mô hình cho từng loại để thực hiện dự đoán xem bộ phim là tích cực, tiêu cực hay trung lập.*

**3.1 Naïve Bayes**

**3.2 Logistic Regression**

**3.3 Support Vector Machine**