ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA KHOA HỌC MÁY TÍNH

BÁO CÁO ĐỒ ÁN
MÁY HỌC

ĐỀ TÀI
NEGATIVE NEWS FILTERING

Giảng viên: Lê Đình Duy
Phạm Nguyễn Trường An
Lớp: CS114.O11.KHCL
Sinh viên thực hiện:
1. Trần Duy Tùng - 21522770
2. Nguyễn Quốc Hưng - 21520253
3. Nguyễn Đinh Minh Chí - 21520648

TP.HCM, ngày 28 tháng 1 năm 2023

TÓM TẮT ĐỒ ÁN

Vì nhận thấy hiện nay những tin tức tiêu cực tràn lan trên khắp mạng xã hội cũng như các trang báo có thể gây ảnh hưởng đến tinh thần cũng như sức khỏe đến những người sử dụng Internet và cũng vì chưa có một ứng dụng nào có thể hạn chế việc xuất hiện nhiều tin tức tiêu cực nên nhóm quyết định thực hiện đồ án "Filtering Negative News".

Nhóm em hướng tới một ứng dụng hoạt động như cảnh báo video bạo lực của Facecbook là sẽ làm mờ những tin được cho là 'tiêu cực' và sẽ đưa ra cảnh báo trước khi một người nào đó muốn đọc. Việc làm vậy sẽ làm hạn chế những tin tức hoặc bài đăng có những thông tin mà chúng ta không muốn tiếp nhận có thể tạo ra một môi trường thông tin lành mạnh trên Internet.

Nhóm em sẽ crawl dữ liệu là các bài báo online bao gồm tiêu đề và nội dung và tự gán nhãn. Sau khi xây dựng bộ dữ liệu hoàn chỉnh, nhóm sẽ tiền xử lí dữ liệu và cuối cùng sẽ dùng các mô hình học máy để thực hiện đồ án này.

Input

– Tiêu đề kết hợp với nội dung trong mỗi bài báo.

Output

– Dự đoán bài báo đó là 'Negative' hay 'Non-Negative'.

Link github: https://github.com/NQHung1/CS114.O11.KHCL

LỜI CẢM ƠN

Nhóm em xin trân trọng cảm ơn thầy Phạm Nguyễn Trường An và thầy Lê Đình Duy đã đồng hành cùng bọn em trong suốt thời gian học môn CS114.

Nhờ có môn học này mà bọn em được thực hành việc tìm kiếm dữ liệu, thu thập cũng như tự gán nhãn dữ liệu thay vì dùng những dữ liệu có sẵn như trước đây. Nhóm em cũng đã hiểu được tầm quan trọng của dữ liệu cũng như việc định nghĩa và mô tả bài toán.

Vì là lần đầu thực hiện dạng đồ án như này nên nhóm đã gặp rất nhiều khó khăn nhưng chúng em cũng đã rất nỗ lực nên trong quá trình viết bài cũng như buổi báo cáo đồ án môn học chúng em còn nhiều điều thiếu sót, kính mong sự góp ý và giúp đỡ của các thầy.

Trân trọng !

CHƯƠNG SỐ 0

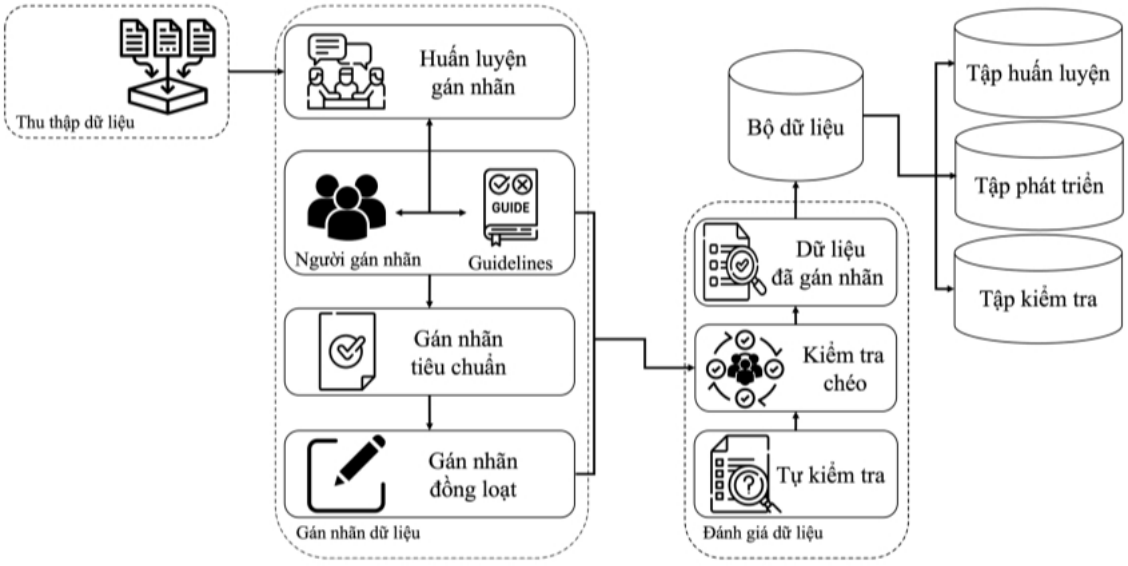
Nhóm em cập nhập lại 2 phần chính:

Định nghĩa rõ lại 'tin tức tiêu cực':

Tin tức tiêu cực là những tin có nội dung phản ánh những sự kiện, hiện tượng, hành vi có thể làm cho người đọc cảm thấy lo lắng, xót xa, thương cảm, đau khổ hoặc chứa những thông tin bạo lực, xung đột, mang tính chất kích động, thù địch gây tổn hại, bất bình cho con người, xã hội. Để giải thích rõ hơn nhóm đã chia ra nhiều lĩnh vực khác nhau bao gồm:

- Thời sự: Các tin tức về tham nhũng, hối lộ, bất ổn chính trị, biểu tình, xung đột và chiến tranh...
- Kinh tế: Các tin tức về suy thoái kinh tế, lạm phát, thất nghiệp, giá cả leo thang,...
- Xã hội: Các tin tức về tội phạm, bạo lực, tai nạn, ô nhiễm môi trường, thiên tai, thảm họa tự nhiên,...
- Y tế: Các tin tức về dịch bệnh, các vấn đề sức khỏe đặc biệt nghiêm trọng hoặc đau lòng, tử vong,...
- Thể thao: Các tin tức về đả kích, chơi xấu, chấn thương,...
- Giáo dục: Các tin tức về vi phạm đạo đức nhà giáo, gian lận trong thi cử, bạo lực học đường,...

Thay đổi quy trình xây dựng bộ dữ liệu:



Quy trình xây dựng bộ dữ liệu.

1. Thu thập dữ liệu:

Sử dụng thư viện Scrapy để thu thập dữ liệu các bài báo trên các trang báo online. Nhóm quyết định chọn một trong những trang báo được truy cập nhiều nhất là Vn-Express.

Tuần tự các bước thực hiện:

- Cài đặt thư viện scrapy: *pip install scrapy*
- Tạo project mới: *scrapy startproject tutorial*
- Tạo 1 file *Spider.py* - một đối tượng Python được xây dựng để tự động hóa quá trình trích xuất dữ liệu từ trang web cụ thể.
- Cuối cùng thực hiện lệnh để lấy dữ liệu về *!scrapy crawl news -o dataset.json*

2. Huấn luyện gán nhãn và gán nhãn:

Nhóm sẽ sử dụng định nghĩa 'tin tức tiêu cực' làm guidlines và cũng như các các tin liên quan đã liệt kê ở từng lĩnh vực để phân loại tin nào là tin tiêu cực.

Sau đó, nhóm đã có một buổi meeting để gán nhãn tiêu chuẩn với 50 tin bất kì ứng với từng linh vực để mỗi thành viên có sự đồng đều và gán sao cho thật chính xác.

Cuối cùng là thực hiện việc gán nhãn đồng loạt. Để tăng cường độ chính xác của dữ liệu thì mỗi thành viên sẽ gán nhãn toàn bộ dữ liệu và sau đó sẽ thực hiện việc kiểm tra chéo.

HƯNG	TÙNG	CHÍ	title	text
0	1	1	Triệu chứng zona thần kinh trên da	Zona thần kinh còn gọi là zona hoặc shingles, là bệnh nhiễm trùng da do virus Varicella zoster (VZ
1	0	0	Ăn thịt chó có hại?	Trời chuyển lạnh, thịt chó và rượu trở thành món nhậu ưa thích của nhiều quý ông bởi hương vị
0	1	1	Lý do ít ngờ gây khó thụ thai	Các chuyên gia y tế khuyến cáo không nên thụt rửa bộ phận sinh dục nhiều vì có thể nhiễm trùng
0	1	0	Giải mã hiện tượng ốm nghén	Nghiên cứu được xuất bản trên tạp chí Nature, số ra tháng 12/2023. Theo các chuyên gia, phụ nữ
0	1	1	Sai lầm khi ăn sáng dễ làm tăng đường huyết	Bữa sáng hỗ trợ người bệnh tiểu đường kiểm soát đường huyết, ảnh hưởng đến tâm trạng, năng
0	1	1	Dấu hiệu mắc bệnh lậu	Theo Trung tâm Kiểm soát và Phòng ngừa Bệnh tật Mỹ (CDC), bệnh lậu là bệnh lây truyền qua đư
0	1	1	Khàn tiếng cảnh báo bệnh gì?	Khàn giọng là sự thay đổi về chất lượng hoặc cao độ của giọng nói, khiến âm thanh trở nên thô, k
0	1	1	Các loài hoa ngày Tết đẹp nhưng có độc	Ngày 11/1, bác sĩ Huỳnh Tấn Vũ, Bệnh viện Đại học Y Dược TP HCM - cơ sở 3, cho biết số cây cảnh
0	1	1	Những bệnh phụ nữ dễ mắc hơn nam giới	Ngoài nguy cơ mắc bệnh phụ nữ như ung thư cổ tử cung, hội chứng buồng trứng đa nang và tiền
0	1	1	Sai lầm khi ăn uống gây rụng tóc	Rụng 100 sợi tóc mỗi ngày là điều bình thường nhưng nhiều hơn có thể do chế độ ăn uống chưa
0	1	1	Các dạng cong vẹo cột sống thường gặp	Vẹo cột sống là hiện tượng cột sống bị cong sang một bên, làm cho xương sườn hoặc cơ nhô ra x
0	1	1	Quan hệ tình dục bằng miệng có lây virus HPV?	HPV là virus gây u nhú ở người, có thể lây qua quan hệ tình dục, khi mẹ sinh em bé và qua các vậ
0	1	1	Dấu hiệu cảnh báo đường huyết cao	Glucose (đường) là chất dinh dưỡng quan trọng nhất cung cấp năng lượng cho cơ thể. Thông thu
0	1	1	Điện thoại sạc qua đêm phát nổ khiến cặp vợ chồng bỏng nặng	Ngày 28/12, đại diện Bệnh viện Đa khoa Hùng Vương cho biết hai bệnh nhân được đưa vào cấp c
0	1	1	6 thói quen thường gặp dễ làm tăng đường huyết	ThS.BSCKI Hà Thị Ngọc Bích, khoa Nội tiết - Đái tháo đường, Bệnh viện Đa khoa Tâm Anh TP HCM
0	1	1	6 bệnh ở đại tràng gây đau bụng	Đại tràng có chiều dài khoảng 1-2 m, thực hiện chức năng co bóp để tống chất thải, thức ăn đã ti
0	1	1	Thủ phạm gây bệnh bạch cầu	Bệnh bạch cầu là loại ung thư máu, do một loạt đột biến ở gene kiểm soát sự phát triển của tế bà
0	1	1	Chứng mất mùi	- Viêm nhiễm: Viêm mũi xoang, polyp mũi... thường hồi phục sau điều trị viêm nhiễm.- Viêm màn
0	1	1	Thủ phạm khiến cholesterol cao	Các thói quen lành mạnh như ăn uống cân bằng, tập thể dục thường xuyên, kiểm soát cân nặng c
0	1	1	Bệnh buồng trứng đa nang	Bác sĩ Nguyễn Minh Thúy, Trung tâm Hỗ trợ Sinh sản, bệnh viện Đa khoa Tâm Anh Hà Nội (IVFTA),

Gán nhãn dữ liệu

3. Kiểm tra chéo và hoàn tất dữ liệu:

Sau khi hoàn tất việc gán nhãn đồng loạt nhóm đã thống kê những dữ liệu khác nhau về việc gán nhãn. Tiếp theo, nhóm sẽ tiếp tục meeting và sẽ cùng nhau phân loại tất cả những nhãn không đồng nhất. Cuối cùng sẽ có được bộ dữ liệu hoàn chỉnh.

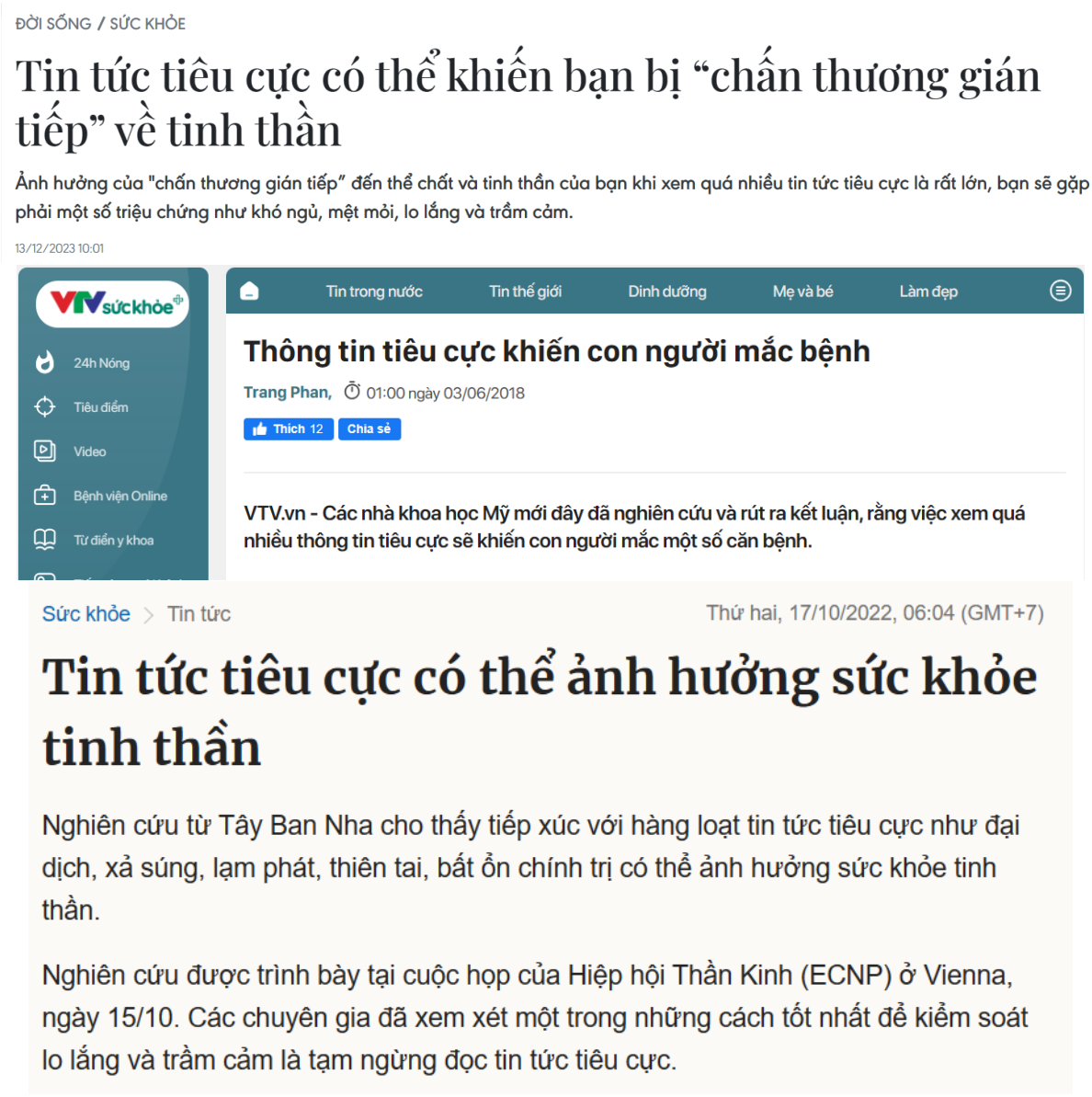
Final	HƯNG	TÙNG	CHÍ	title	text
1	0	1	1	Triệu chứng zona thần kinh trên da	Zona thần kinh còn gọi là zona hoặc shingles, là bệnh nhiễm trùng da do virus Varicella zoster (VZ
1	1	0	0	Ăn thịt chó có hại?	Trời chuyển lạnh, thịt chó và rượu trở thành món nhậu ưa thích của nhiều quý ông bởi hương vị t
1	0	1	1	Lý do ít ngờ gây khó thụ thai	Các chuyên gia y tế khuyến cáo không nên thụt rửa bộ phận sinh dục nhiều vì có thể nhiễm trùng
0	0	1	0	Giải mã hiện tượng ốm nghén	Nghiên cứu được xuất bản trên tạp chí Nature, số ra tháng 12/2023. Theo các chuyên gia, phụ nữ
1	0	1	1	Sai lầm khi ăn sáng dễ làm tăng đường huyết	Bữa sáng hỗ trợ người bệnh tiểu đường kiểm soát đường huyết, ảnh hưởng đến tâm trạng, năng
1	0	1	1	Dấu hiệu mắc bệnh lậu	Theo Trung tâm Kiểm soát và Phòng ngừa Bệnh tật Mỹ (CDC), bệnh lậu là bệnh lây truyền qua đư
1	0	1	1	Khàn tiếng cảnh báo bệnh gì?	Khàn giọng là sự thay đổi về chất lượng hoặc cao độ của giọng nói, khiến âm thanh trở nên thô, k
1	0	1	1	Các loài hoa ngày Tết đẹp nhưng có độc	Ngày 11/1, bác sĩ Huỳnh Tấn Vũ, Bệnh viện Đại học Y Dược TP HCM - cơ sở 3, cho biết số cây cảnh
1	0	1	1	Những bệnh phụ nữ dễ mắc hơn nam giới	Ngoài nguy cơ mắc bệnh phụ nữ như ung thư cổ tử cung, hội chứng buồng trứng đa nang và tiền
1	0	1	1	Sai lầm khi ăn uống gây rụng tóc	Rụng 100 sợi tóc mỗi ngày là điều bình thường nhưng nhiều hơn có thể do chế độ ăn uống chưa
1	0	1	1	Các dạng cong vẹo cột sống thường gặp	Vẹo cột sống là hiện tượng cột sống bị cong sang một bên, làm cho xương sườn hoặc cơ nhô ra x
1	0	1	1	Quan hệ tình dục bằng miệng có lây virus HPV?	HPV là virus gây u nhú ở người, có thể lây qua quan hệ tình dục, khi mẹ sinh em bé và qua các vật
1	0	1	1	Dấu hiệu cảnh báo đường huyết cao	Glucose (đường) là chất dinh dưỡng quan trọng nhất cung cấp năng lượng cho cơ thể. Thông thư
1	0	1	1	Điện thoại sạc qua đêm phát nổ khiến cặp vợ chồng bỏng nặng	Ngày 28/12, đại diện Bệnh viện Đa khoa Hùng Vương cho biết hai bệnh nhân được đưa vào cấp c
1	0	1	1	6 thói quen thường gặp dễ làm tăng đường huyết	ThS.BSCKI Hà Thị Ngọc Bích, khoa Nội tiết - Đái tháo đường, Bệnh viện Đa khoa Tâm Anh TP HCM
1	0	1	1	6 bệnh ở đại tràng gây đau bụng	Đại tràng có chiều dài khoảng 1-2 m, thực hiện chức năng co bóp để tống chất thải, thức ăn đã tiê
1	0	1	1	Thủ phạm gây bệnh bạch cầu	Bệnh bạch cầu là loại ung thư máu, do một loạt đột biến ở gene kiểm soát sự phát triển của tế bà
0	0	1	1	Chứng mất mùi	- Viêm nhiễm: Viêm mũi xoang, polyp mũi... thường hồi phục sau điều trị viêm nhiễm.- Viêm màn
0	0	1	1	Thủ phạm khiến cholesterol cao	Các thói quen lành mạnh như ăn uống cân bằng, tập thể dục thường xuyên, kiểm soát cân nặng c
1	0	1	1	Bệnh buồng trứng đa nang	Bác sĩ Nguyễn Minh Thúy, Trung tâm Hỗ trợ Sinh sản, bệnh viện Đa khoa Tâm Anh Hà Nội (IVFTA),

Kiểm tra dữ liệu và thống nhất

MỤC LỤC

1 Tổng quan

1.1 Động lực thực hiện



Tác hại 'tin tức tiêu cực'.

Trong thời kì bùng nổ về Internet và các trang mạng xã hội cũng như các trang báo chí trực tuyến, mọi người có thể dễ dàng tiếp cận thông tin, tin tức một cách tiện lợi và nhanh chóng. Tuy nhiên, đi kèm với sự dễ dàng đó không ít thông tin tiêu cực, sai lệch hoặc gây hại được tạo ra và lan truyền với mục đích thu hút sự chú ý, câu like, câu view, hoặc để hạ uy tín của cá nhân, tổ chức.

Tiếp xúc quá nhiều với tin tiêu cực có thể gây ra những tác động tiêu cực đến cuộc sống của con người, bao gồm:

- Ảnh hưởng đến tâm lý: Tin tiêu cực thường mang tính chất kích động, tiêu cực,

khiến cho người tiếp xúc trở nên bi quan, chán nản, lo lắng, thậm chí là trầm cảm.

- Ảnh hưởng đến sức khỏe: Tin tiêu cực có thể gây ra những rối loạn về giấc ngủ, ăn uống, thậm chí là các bệnh lý về tim mạch, huyết áp.
- Ảnh hưởng đến nhận thức: Tin tiêu cực có thể khiến cho người tiếp xúc có cái nhìn sai lệch về thế giới, về xã hội, về con người.
- Ảnh hưởng đến hành vi: Tin tiêu cực có thể kích động, lôi kéo người tiếp xúc thực hiện những hành vi tiêu cực, thậm chí là phạm pháp.

Nhóm em nhận thấy bài toán 'Filtering negative news' có thể giúp người đọc tin tránh được những thông tin có thể gây hại đến tinh thần cũng như sức khỏe của người sử dụng Internet, góp phần vào việc tạo ra một môi trường thông tin lành mạnh.

1.2 Mô tả bài toán

Input

– Tiêu đề kết hợp với nội dung trong mỗi bài báo.

Output

– Dự đoán bài báo đó là 'Negative' hay 'Non-Negative'.

Định nghĩa 'tin tức tiêu cực'

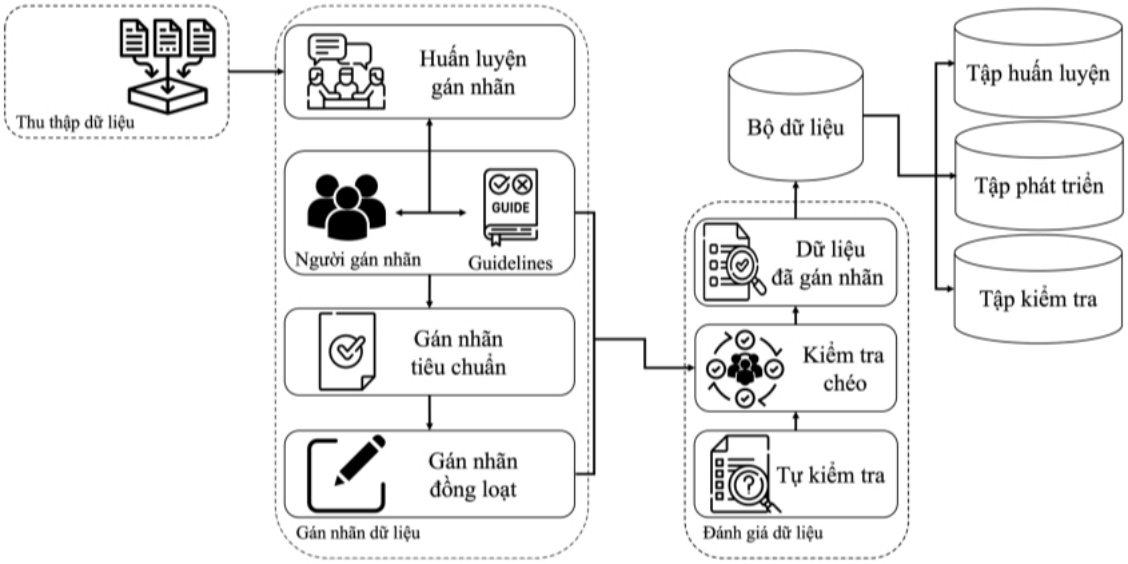
Nhóm em định nghĩa tin tức tiêu cực là những tin có nội dung bi quan, chứa những thông tin bạo lực, xung đột, hay thông tin mang tính chất kích động, thù địch. Để giải thích rõ hơn nhóm đã chia ra nhiều lĩnh vực khác nhau bao gồm:

- Thời sự: Các tin tức về tham nhũng, hối lộ, bất ổn chính trị, biểu tình, xung đột và chiến tranh...
- Kinh tế: Các tin tức về suy thoái kinh tế, lạm phát, thất nghiệp, giá cả leo thang,...
- Xã hội: Các tin tức về tội phạm, bạo lực, tai nạn, ô nhiễm môi trường, thiên tai, thảm họa tự nhiên,...
- Y tế: Các tin tức về dịch bệnh, các vấn đề sức khỏe đặc biệt nghiêm trọng hoặc đau lòng, tử vong,...
- Thể thao: Các tin tức về đả kích, chơi xấu, chấn thương,...
- Giáo dục: Các tin tức về vi phạm đạo đức nhà giáo, gian lận trong thi cử, bạo lực học đường,...

2 Xây dựng bộ dữ liệu

2.1 Quy trình xây dựng

Nhóm sẽ thực hiện việc xây dựng bộ dữ liệu theo quy trình như sau:



Quy trình xây dựng bộ dữ liệu.

1. **Thu thập dữ liệu:** Sử dụng thư viện Scrapy để thu thập dữ liệu các bài báo trên các trang báo online. Nhóm quyết định chọn một trong những trang báo được truy cập nhiều nhất là VnExpress.

   Tuần tự các bước thực hiện:

   – Cài đặt thư viện scrapy: *pip install scrapy*

   – Tạo project mới: *scrapy startproject tutorial*

   – Tạo 1 file *Spider.py* - một đối tượng Python được xây dựng để tự động hóa quá trình trích xuất dữ liệu từ trang web cụ thể.

   – Cuối cùng thực hiện lệnh để lấy dữ liệu về *!scrapy crawl news -o dataset.json*

```python
import scrapy

class NewsSpider(scrapy.Spider):
    name = "news"

    def start_requests(self):
        urls = [
            'https://vnexpress.net/phap-luat'
        ]
        for url in urls:
            yield scrapy.Request(url, callback=self.parse)

    def parse(self, response):
        allLinks = response.css('h3.title-news a::attr(href)').getall()

        for each_link in allLinks:
            yield scrapy.Request(each_link, callback=self.parse_href)

        next_page = response.css('a.btn-page next-page::attr(href)').get()
        if next_page is not None:
            next_page = response.urljoin(next_page)
            yield scrapy.Request(next_page, callback=self.parse)

    def parse_href(self, response):
        title = response.css('.title-detail::text').get()
        title = title.replace('\n', '')
        title = title.replace('\t', '')
        paragraph = response.css('.fck_detail p::text').getall()
        text = ''
        count = 0
        for sentence in paragraph:
            if count == 4:
                break
            sentence = sentence.replace('\xa0', '')
            text = text + sentence
            count += 1
        yield {
            'title': title,
            'text': text,
        }
```

Spider.py

2. Huấn luyện gán nhãn và gán nhãn:

Nhóm sẽ sử dụng định nghĩa ’tin tức tiêu cực’ làm guidlines và cũng như các các tin liên quan đã liệt kê ở từng lĩnh vực để phân loại tin nào là tin tiêu cực.

Sau đó, nhóm đã có một buổi meeting để gán nhãn tiêu chuẩn với 50 tin bất kì ứng với từng linh vực để mỗi thành viên có sự đồng đều và gán sao cho thật chính xác.

Cuối cùng là thực hiện việc gán nhãn đồng loạt. Để tăng cường độ chính xác của dữ liệu thì mỗi thành viên sẽ gán nhãn toàn bộ dữ liệu và sau đó sẽ thực hiện việc

kiểm tra chéo.

HƯNG	TÙNG	CHÍ	title	text
0	1	1	Triệu chứng zona thần kinh trên da	Zona thần kinh còn gọi là zona hoặc shingles, là bệnh nhiễm trùng da do virus Varicella zoster (VZ
1	0	0	Ăn thịt chó có hại?	Trời chuyển lạnh, thịt chó và rượu trở thành món nhậu ưa thích của nhiều quý ông bởi hương vị
0	1	1	Lý do ít ngờ gây khó thụ thai	Các chuyên gia y tế khuyến cáo không nên thụt rửa bộ phận sinh dục nhiều vì có thể nhiễm trùng
0	1	0	Giải mã hiện tượng ốm nghén	Nghiên cứu được xuất bản trên tạp chí Nature, số ra tháng 12/2023. Theo các chuyên gia, phụ nữ
0	1	1	Sai lầm khi ăn sáng dễ làm tăng đường huyết	Bữa sáng hỗ trợ người bệnh tiểu đường kiểm soát đường huyết, ảnh hưởng đến tâm trạng, năng
0	1	1	Dấu hiệu mắc bệnh lậu	Theo Trung tâm Kiểm soát và Phòng ngừa Bệnh tật Mỹ (CDC), bệnh lậu là bệnh lây truyền qua đư
0	1	1	Khàn tiếng cảnh báo bệnh gì?	Khàn giọng là sự thay đổi về chất lượng hoặc cao độ của giọng nói, khiến âm thanh trở nên thô, k
0	1	1	Các loài hoa ngày Tết đẹp nhưng có độc	Ngày 11/1, bác sĩ Huỳnh Tấn Vũ, Bệnh viện Đại học Y Dược TP HCM - cơ sở 3, cho biết số cây cảnh
0	1	1	Những bệnh phụ nữ dễ mắc hơn nam giới	Ngoài nguy cơ mắc bệnh phụ nữ như ung thư cổ tử cung, hội chứng buồng trứng đa nang và tiền
0	1	1	Sai lầm khi ăn uống gây rụng tóc	Rụng 100 sợi tóc mỗi ngày là điều bình thường nhưng nhiều hơn có thể do chế độ ăn uống chưa
0	1	1	Các dạng cong vẹo cột sống thường gặp	Vẹo cột sống là hiện tượng cột sống bị cong sang một bên, làm cho xương sườn hoặc cơ nhô ra x
0	1	1	Quan hệ tình dục bằng miệng có lây virus HPV?	HPV là virus gây u nhú ở người, có thể lây qua quan hệ tình dục, khi mẹ sinh em bé và qua các vậ
0	1	1	Dấu hiệu cảnh báo đường huyết cao	Glucose (đường) là chất dinh dưỡng quan trọng nhất cung cấp năng lượng cho cơ thể. Thông thu
0	1	1	Điện thoại sạc qua đêm phát nổ khiến cặp vợ chồng bỏng nặng	Ngày 28/12, đại diện Bệnh viện Đa khoa Hùng Vương cho biết hai bệnh nhân được đưa vào cấp c
0	1	1	6 thói quen thường gặp dễ làm tăng đường huyết	ThS.BSCKI Hà Thị Ngọc Bích, khoa Nội tiết - Đái tháo đường, Bệnh viện Đa khoa Tâm Anh TP HCM
0	1	1	6 bệnh ở đại tràng gây đau bụng	Đại tràng có chiều dài khoảng 1-2 m, thực hiện chức năng co bóp để tống chất thải, thức ăn đã tiê
0	1	1	Thủ phạm gây bệnh bạch cầu	Bệnh bạch cầu là loại ung thư máu, do một loạt đột biến ở gene kiểm soát sự phát triển của tế bà
0	1	1	Chứng mất mùi	- Viêm nhiễm: Viêm mũi xoang, polyp mũi... thường hồi phục sau điều trị viêm nhiễm.- Viêm màn
0	1	1	Thủ phạm khiến cholesterol cao	Các thói quen lành mạnh như ăn uống cân bằng, tập thể dục thường xuyên, kiểm soát cân nặng c
0	1	1	Bệnh buồng trứng đa nang	Bác sĩ Nguyễn Minh Thúy, Trung tâm Hỗ trợ Sinh sản, bệnh viện Đa khoa Tâm Anh Hà Nội (IVFTA),

Gán nhãn dữ liệu

3. Kiểm tra chéo và hoàn tất dữ liệu:

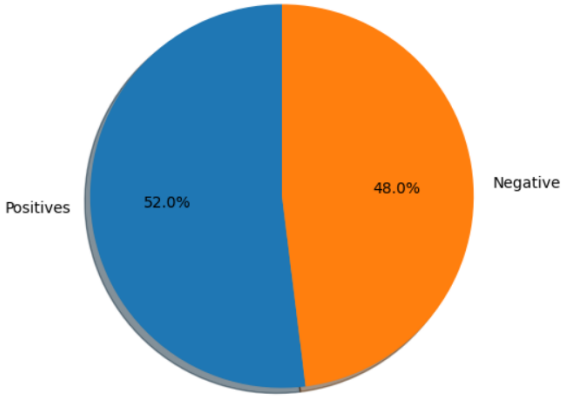
Sau khi hoàn tất việc gán nhãn đồng loạt nhóm đã thống kê những dữ liệu khác nhau về việc gán nhãn. Tiếp theo, nhóm sẽ tiếp tục meeting và sẽ cùng nhau phân loại tất cả những nhãn không đồng nhất. Cuối cùng sẽ có được bộ dữ liệu hoàn chỉnh.

Final	HƯNG	TÙNG	CHÍ	title	text
1	0	1	1	Triệu chứng zona thần kinh trên da	Zona thần kinh còn gọi là zona hoặc shingles, là bệnh nhiễm trùng da do virus Varicella zoster (VZ
1	1	0	0	Ăn thịt chó có hại?	Trời chuyển lạnh, thịt chó và rượu trở thành món nhậu ưa thích của nhiều quý ông bởi hương vị t
1	0	1	1	Lý do ít ngờ gây khó thụ thai	Các chuyên gia y tế khuyến cáo không nên thụt rửa bộ phận sinh dục nhiều vì có thể nhiễm trùng
0	0	1	0	Giải mã hiện tượng ốm nghén	Nghiên cứu được xuất bản trên tạp chí Nature, số ra tháng 12/2023. Theo các chuyên gia, phụ nữ
1	0	1	1	Sai lầm khi ăn sáng dễ làm tăng đường huyết	Bữa sáng hỗ trợ người bệnh tiểu đường kiểm soát đường huyết, ảnh hưởng đến tâm trạng, năng
1	0	1	1	Dấu hiệu mắc bệnh lậu	Theo Trung tâm Kiểm soát và Phòng ngừa Bệnh tật Mỹ (CDC), bệnh lậu là bệnh lây truyền qua đư
1	0	1	1	Khàn tiếng cảnh báo bệnh gì?	Khàn giọng là sự thay đổi về chất lượng hoặc cao độ của giọng nói, khiến âm thanh trở nên thô, k
1	0	1	1	Các loài hoa ngày Tết đẹp nhưng có độc	Ngày 11/1, bác sĩ Huỳnh Tấn Vũ, Bệnh viện Đại học Y Dược TP HCM - cơ sở 3, cho biết số cây cảnh
1	0	1	1	Những bệnh phụ nữ dễ mắc hơn nam giới	Ngoài nguy cơ mắc bệnh phụ nữ như ung thư cổ tử cung, hội chứng buồng trứng đa nang và tiền
1	0	1	1	Sai lầm khi ăn uống gây rụng tóc	Rụng 100 sợi tóc mỗi ngày là điều bình thường nhưng nhiều hơn có thể do chế độ ăn uống chưa
1	0	1	1	Các dạng cong vẹo cột sống thường gặp	Vẹo cột sống là hiện tượng cột sống bị cong sang một bên, làm cho xương sườn hoặc cơ nhô ra x
1	0	1	1	Quan hệ tình dục bằng miệng có lây virus HPV?	HPV là virus gây u nhú ở người, có thể lây qua quan hệ tình dục, khi mẹ sinh em bé và qua các vật
1	0	1	1	Dấu hiệu cảnh báo đường huyết cao	Glucose (đường) là chất dinh dưỡng quan trọng nhất cung cấp năng lượng cho cơ thể. Thông thư
1	0	1	1	Điện thoại sạc qua đêm phát nổ khiến cặp vợ chồng bỏng nặng	Ngày 28/12, đại diện Bệnh viện Đa khoa Hùng Vương cho biết hai bệnh nhân được đưa vào cấp c
1	0	1	1	6 thói quen thường gặp dễ làm tăng đường huyết	ThS.BSCKI Hà Thị Ngọc Bích, khoa Nội tiết - Đái tháo đường, Bệnh viện Đa khoa Tâm Anh TP HCM
1	0	1	1	6 bệnh ở đại tràng gây đau bụng	Đại tràng có chiều dài khoảng 1-2 m, thực hiện chức năng co bóp để tống chất thải, thức ăn đã tiê
1	0	1	1	Thủ phạm gây bệnh bạch cầu	Bệnh bạch cầu là loại ung thư máu, do một loạt đột biến ở gene kiểm soát sự phát triển của tế bà
0	0	1	1	Chứng mất mùi	- Viêm nhiễm: Viêm mũi xoang, polyp mũi... thường hồi phục sau điều trị viêm nhiễm.- Viêm màn
0	0	1	1	Thủ phạm khiến cholesterol cao	Các thói quen lành mạnh như ăn uống cân bằng, tập thể dục thường xuyên, kiểm soát cân nặng c
1	0	1	1	Bệnh buồng trứng đa nang	Bác sĩ Nguyễn Minh Thúy, Trung tâm Hỗ trợ Sinh sản, bệnh viện Đa khoa Tâm Anh Hà Nội (IVFTA),

Kiểm tra dữ liệu và thống nhất

2.2 Một tả dữ liệu

Bộ dữ liệu tổng cộng 4499 news thuộc 6 chủ đề khác nhau trong đó gồm 2161 'Negative news' và 2338 'Non-negative news'.



Thống kê dữ liệu

2.3 Xử lí dữ liệu

– Loại bỏ các giá trị 'NaN'.

```
#Remove NaN
if dataset.isnull().values.any():
  dataset = dataset.dropna()
dataset
```

– Loại bỏ các giá trị lặp lại.

```
#Remove Duplication
if dataset['Title'].duplicated().any():
  dataset = dataset.drop_duplicates()
dataset.shape
```

– Đưa tất cả tiêu đề và nội dung về dạng chữ thường.

```
#Lowercasing
dataset[['Title','Text']] = dataset[['Title','Text']].apply(lambda x: x.str.lower())
dataset.head()
```

– Loại bỏ các dấu câu.

```
#Remove Punctuations
def remove_punctuation(title):
    title = [char for char in title if char not in string.punctuation]
    removed = ''.join(title)
    return removed
dataset['Title'] = dataset['Title'].apply(remove_punctuation)
dataset['Text'] = dataset['Text'].apply(remove_punctuation)
dataset
```

– 'Tokenizer' chia nội dung thành các cụm từ có nghĩa, hiệu quả cho việc huấn luyện mô hình. Nhóm em sử dụng thư viện 'pyvi' đã được huấn luyện trên bộ dữ liệu tiếng việt rất lớn.

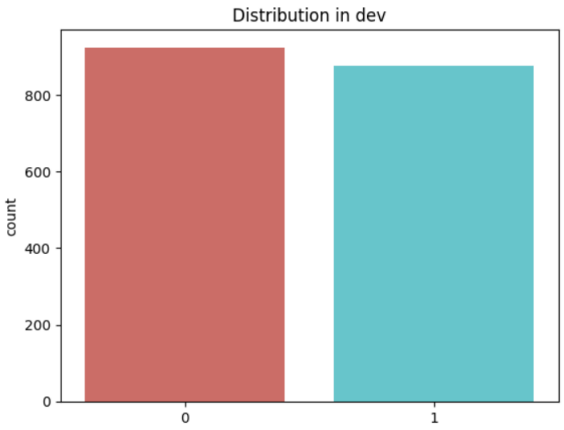
```
#Tokenizer
dataset['Title'] = dataset['Title'].apply(ViTokenizer.tokenize)
dataset['Text'] = dataset['Text'].apply(ViTokenizer.tokenize)
dataset.head()
```

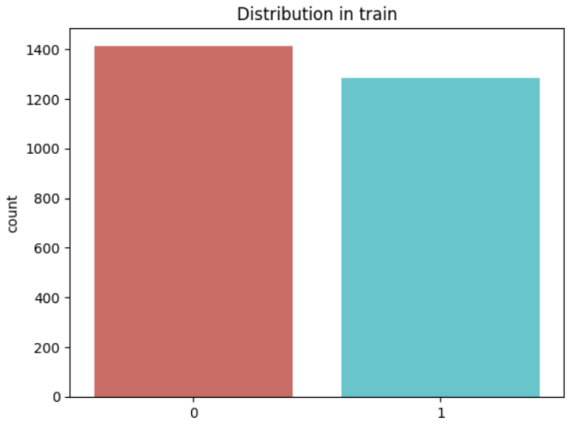
– Loại bỏ các stopwords. Nhóm đã tìm hiểu và sử dụng bộ 'Vietnamese-stopwords' tải về từ github: vietnamese-stopwords

```
import requests

# URL of file stopwords
url = 'https://raw.githubusercontent.com/stopwords/vietnamese-stopwords/master/vietnamese-stopwords.txt'

# Request
response = requests.get(url)

if response.status_code == 200:
    # Save file 'vietnamese-stopwords.txt'
    with open('vietnamese-stopwords.txt', 'w') as f:
        f.write(response.text)
else:
    print('Error:', response.status_code)

with open('vietnamese-stopwords.txt', 'r') as f:
    vietnamese_stopwords = f.read().splitlines()
```

2.4 Chia dữ liệu huấn luyện

Chia bộ dữ liệu thành tập train và dev theo tỉ lệ 60:40 trong đó tập train gồm 2699 news và tập dev gồm 1800 news.





3 Huấn luyện và đánh giá các mô hình

3.1 Trích xuất đặc trưng

– Bag of Words (BOW) là một phương pháp để trích xuất các đặc điểm từ các dữ liệu văn bản. Nó tạo ra một kho từ vựng chứa tất cả các từ duy nhất có trong tất cả các dữ liệu văn bản trong tập huấn luyện.

– Term frequency – inverse document frequency(Tf-idf) là trọng số của một từ trong văn bản thu được qua thống kê thể hiện mức độ quan trọng của từ này trong một văn bản, mà bản thân văn bản đang xét nằm trong một tập hợp các văn bản.

3.2 Model

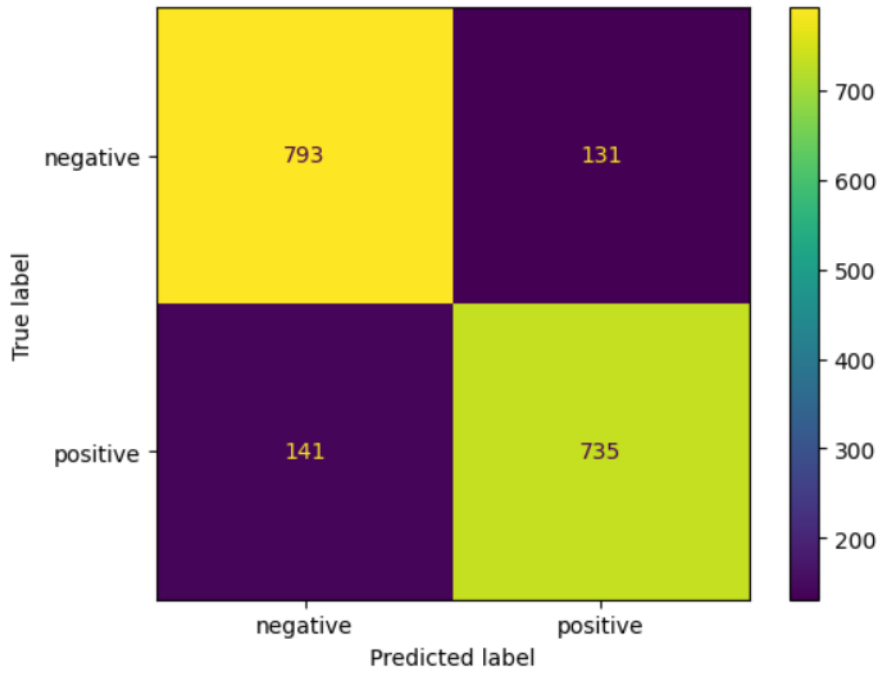
```
class Model:
    def __init__(self, Vectorizer, Model, train, train_is_neg, dev, dev_is_neg):
        self.vectorizer = Vectorizer
        self.model = Model
        self.train = train['Title'] + ' ' + train['Text']
        self.train_is_neg = train_is_neg
        self.dev = dev['Title'] + ' ' + dev['Text']
        self.dev_is_neg = dev_is_neg

    def fit(self):
        self.vectorizer.fit(self.train)
        features = self.vectorizer.transform(self.train)
        self.model.fit(features, self.train_is_neg)

    def evaluate(self):
        features = self.vectorizer.transform(self.dev)
        dev_pred = self.model.predict(features)

        acc = accuracy_score(self.dev_is_neg, dev_pred)
        recall = recall_score(self.dev_is_neg, dev_pred)
        precision = precision_score(self.dev_is_neg, dev_pred)
        f1 = f1_score(self.dev_is_neg, dev_pred)

        print(f"Accuracy: {acc: .4f}")
        print(f"Recall: {recall: .4f}")
        print(f"Precision: {precision: .4f}")
        print(f"F1: {f1: .4f}")
```

Hình: Model

– Logistic Regression là một kỹ thuật phân tích dữ liệu sử dụng toán học để tìm ra mối quan hệ giữa hai yếu tố dữ liệu. Sau đó, kỹ thuật này sử dụng mối quan hệ đã tìm được để dự đoán giá trị của những yếu tố đó dựa trên yếu tố còn lại. Dự đoán thường cho ra một số kết quả hữu hạn, như có hoặc không.

```
model = Model(Count_Vectorizer, LogisticRegression(max_iter = 10000), train, train['Is_neg'], dev, dev['Is_neg'])
model.fit()
model.evaluate()
```

```
Accuracy:  0.8489
Recall:  0.8390
Precision:  0.8487
F1:  0.8439
```

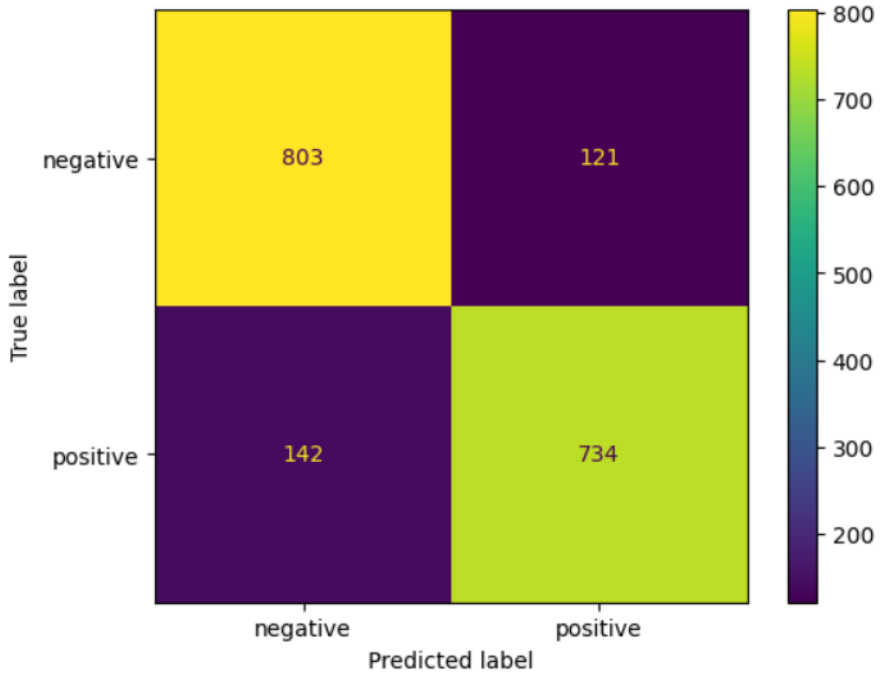


Hình: Model Losistic sử dụng phương pháp BOW

– Support Vectoer Machine là một thuật toán giám sát, nó có thể sử dụng cho cả việc phân loại hoặc đệ quy. Tuy nhiên nó được sử dụng chủ yếu cho việc phân loại. Trong thuật toán này, chúng ta vẽ đồi thị dữ liệu là các điểm trong n chiều (ở đây n là số lượng các tính năng bạn có) với giá trị của mỗi tính năng sẽ là một phần liên kết. Sau đó chúng ta thực hiện tìm "đường bay" (hyper-plane) phân chia các lớp. Hyper-plane nó chỉ hiểu đơn giản là 1 đường thẳng có thể phân chia các lớp ra thành hai phần riêng biệt.

```
model = Model(Count_Vectorizer, svm.SVC(max_iter = 10000), train, train['Is_neg'], dev, dev['Is_neg'])
model.fit()
model.evaluate()
```

```
Accuracy:  0.8539
Recall:  0.8379
Precision:  0.8585
F1:  0.8481
```

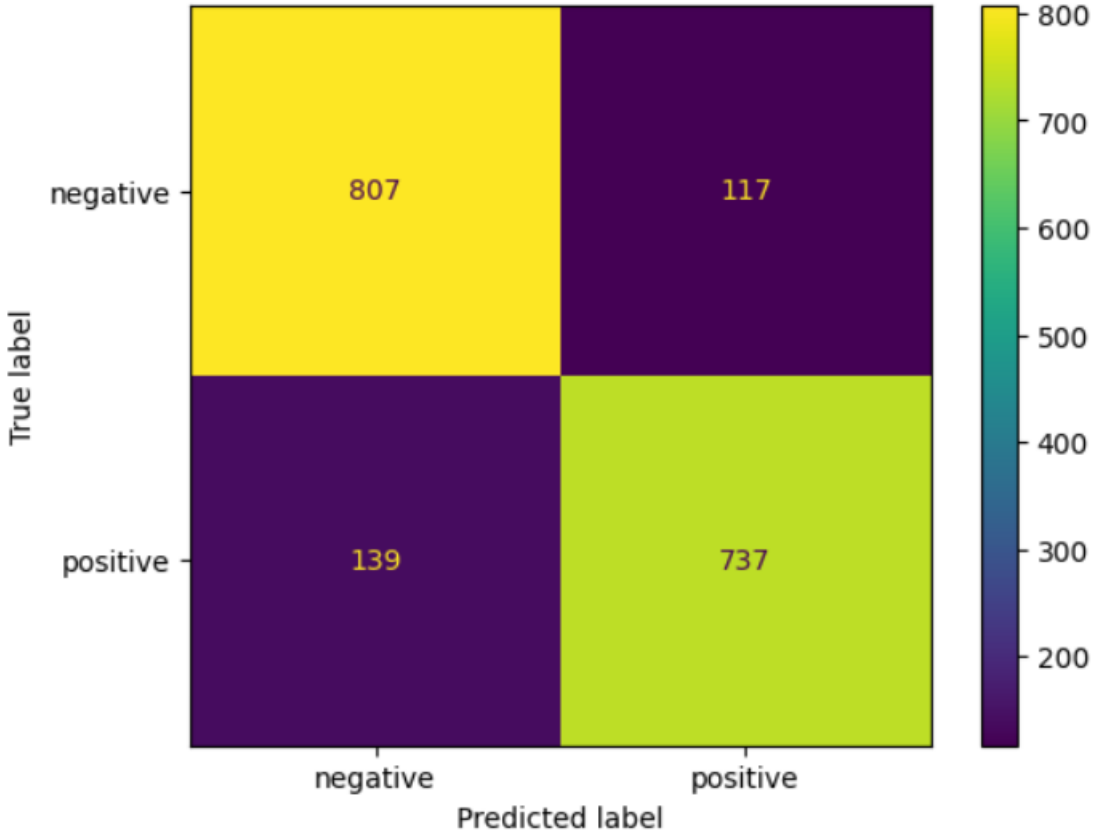


Hình: Model SVM sử dụng phương pháp BOW

– Bộ phân lớp Bayes là một giải thuật thuộc lớp giải thuật thống kê, nó có thể dự đoán xác suất của một phần tử dữ liệu thuộc vào một lớp là bao nhiêu. Phân lớp Bayes được dựa trên định lý Bayes (định lý được đặt theo tên tác giả của nó là Thomas Bayes)

```
model = Model(Count_Vectorizer, BernoulliNB(), train, train['Is_neg'], dev, dev['Is_neg'])
model.fit()
model.evaluate()
```

```
Accuracy:  0.8578
Recall:  0.8413
Precision:  0.8630
F1:  0.8520
```



Hình: Model NB sử dụng phương pháp BOW

3.3 Kết quả

Kết quả sau khi sử dụng 3 model với mỗi model dùng 2 cách trích xuất đặc trưng BOW và Tf-idf

	BOW				Tf-IDF			
	Accuracy	Recall	Precision	F1	Accuracy	Recall	Precision	F1
LR	0.8489	0.8390	0.8487	0.8439	0.8178	0.7671	0.8442	0.8038
SVM	0.8539	0.8379	0.8585	0.8481	0.8289	0.7785	0.8568	0.8158
NB	0.8578	0.8413	0.8630	0.8520	0.7867	0.7751	0.7841	0.7796

Hình: Kết quả.

- Bag of Words giúp model có thể dự đoán chính xác hơn Tfidf .
- Model SVM (BOW) có giá trị accuracy tốt nhất nhưng sau đây tụi em chạy lại và sử dụng Ten-fold cross validation

	BOW				Tf-IDF			
	Accuracy	Recall	Precision	F1	Accuracy	Recall	Precision	F1
LR	0.8504	0.8390	0.8487	0.8439	0.8178	0.7671	0.8442	0.8038
SVM	0.8489	0.8379	0.8585	0.8481	0.8289	0.7785	0.8568	0.8158
NB	0.8486	0.8413	0.8630	0.8520	0.7867	0.7751	0.7841	0.7796

Hình: Kết quả sau khi dùng Ten-fold cross validation.

- Model LR (BOW) có giá trị accuracy tốt nhất với phương pháp K-fold cross validation nên sẽ được chọn làm model chính thức cho bài toán

4 Tài liệu tham khảo

Scrapy - thư viện crawl dữ liệu

Vietnamese stopwords - thư viện stopwords

Pyvi - Tokenizer - thư viện tokenize

Logistic Regression là gì

SVM là gì

Naive Bayes là gì