ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN

ĐỀ TÀI

TOPIC CLASSIFICATION

Môn học: Xử lý ngôn ngữ tự nhiên
(CS221.N11.KHCL)

Giảng viên: ThS. Nguyễn Trọng Chỉnh

Thành viên:

Lê Võ Tiến Phát – 19521993
Nguyễn Thành Trung – 19522432

Thành phố Hồ Chí Minh, ngày 24 tháng 2 năm 2023.

Lời cảm ơn

Đầu tiên, chúng em xin gửi lời cảm ơn và lòng biết ơn sâu sắc nhất tới ThS. Nguyễn Trọng Chỉnh – giảng viên phụ trách bộ môn *"Xử lý ngôn ngữ tự nhiên",* trực thuộc Khoa Khoa học Máy Tính, cũng là cố vấn học tập lớp KHCL2019.3, đã trang bị cho em những kiến thức trong môn học này.

Tuy nhiên trong quá trình nghiên cứu đề tài, do kiến thức chuyên ngành còn hạn chế nên chúng em vẫn còn nhiều thiếu sót khi tìm hiểu, đánh giá, trình bày về đề tài. Những góp ý của thầy trong buổi báo cáo là những kinh nghiệm quý báu mà chúng em học hỏi, khắc phục cho quá trình học tập và làm việc sau này.

Xin chân thành cảm ơn Thầy.

Mục lục

I TỔNG QUAN

1. Ngữ cảnh ứng dụng bài toán

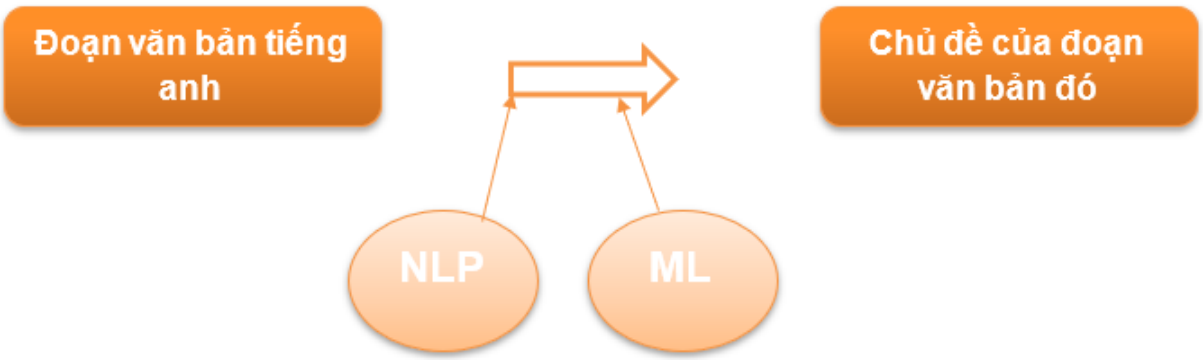
Xử lý ngôn ngữ tự nhiên (NLP) là một lĩnh vực rất quan trọng trong khoa học máy tính và được ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau. Một trong những ngữ cảnh ứng dụng của NLP là trong lĩnh vực tìm kiếm và phân tích thông tin trên web. Với số lượng thông tin khổng lồ trên internet, việc tìm kiếm và phân tích nội dung trở nên cực kỳ khó khăn. Nhưng với NLP, chúng ta có thể xử lý các văn bản, dữ liệu và tìm kiếm nthông ti một cách nhanh chóng và chính xác hơn.

Ngoài ra, NLP còn được ứng dụng trong việc phân tích cảm xúc, tóm tắt văn bản, dịch thuật tự động và phân loại văn bản. Điều này giúp cho người dùng có thể tiết kiệm thời gian và công sức trong công việc hàng ngày, đồng thời giúp các doanh nghiệp nắm bắt được nhu cầu và mong muốn của khách hàng. NLP cũng được sử dụng trong các trình hỗ trợ tương tác giọng nói, giúp cho người dùng có thể tương tác với máy tính bằng giọng nói một cách tự nhiên và dễ dàng hơn.

Đặc biệt là bài toán Topic Classification, đây là bài toán ứng dụng trong lĩnh vực xử lý ngôn ngữ tự nhiên. Nó giúp tổ chức các văn bản thành các chủ đề khác nhau, giúp cho việc tìm kiếm, tra cứu và xử lý dữ liệu trở nên dễ dàng hơn.

2. Bài toán

Như đã nói ở trên, *topic classification* là một bài toán trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) nhằm phân loại các văn bản vào các chủ đề khác nhau dựa trên nội dung của chúng. Nói cách khác, bài toán này giúp cho việc tổ chức, phân loại và đánh giá các văn bản trở nên dễ dàng hơn.



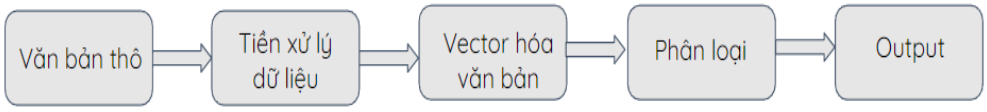
Hình 1 Input và output bài toán

3. Ưu nhược điểm bài toán mang lại

Ưu điểm: Nhìn chung bài toán Topic Classification giúp tổ chức dữ liệu trở nên dễ dàng, đặc biệt là khi có đủ lượng dữ liệu lớn. Phân tích cảm xúc người dùng đối với một chủ đề nhất định và phân loại ra chủ đề phù hợp cho từng người dùng.

Nhược điểm: Để thực hiện được các ưu điểm trên bài toán yêu cầu một lượng dữ liệu vô cùng lớn và đa dạng để huấn luyện mô hình phân loại chính xác.

II NỘI DUNG CHÍNH



Hình 2 Pipeline bài toán

1. Giới thiệu bộ ngữ liệu

Bộ ngữ liệu nhóm sử dụng với ngôn ngữ là tiếng Anh, gồm 60 đoạn văn bản nhỏ (trung bình từ 3 đến 5 câu) chia đều làm 4 chủ đề

Football, Tennis, Golf, Basketball được thu thập từ các trang web: Skysport, CNN, talksport.com, goal.com, nba.com,..

Doc	Categories
Do you remember when you were a kid and you saw that Zinedine Zidane had moved to Real Madrid for £45m? It seemed like so much money at the time - but it's fairly standard these days for chairmen of big clubs across Europe to PayPal each other figures that large for players. £45m these days buys you the right leg of some players. Tens of millions of pounds is still a lot of money in the real world - and yet we baulk at the idea of receiving £20m for a player worth 30, £50m for a player worth 100.	Football
Footage from Barcelona's new Amazon Prime series has revealed the reaction from team-mates after they discovered Lionel Messi was leaving the club in 2021. Last summer, the Argentina forward was forced to walk away from the Camp Nou after his contract came to an end and he was unable to renew due to economic and structural obstacles. It was a stunning moment as the seven-time Ballon d'Or winner left the club to join PSG, and Barcelona's new series shows the shocked reaction from team-mates after they found out the news.	Football
Manchester United were faster, sharper, smarter, and classier than Nottingham Forest in this seventh straight win over their visitors. This was approaching the complete performance: a blend of measured attacking and rapid breaks and the constant harrying of an opponent who ended exhausted and demoralised. Given United had scored only 20 Premier League goals at kickoff and there is limited finance in January to replace Cristiano Ronaldo, a strike apiece from Marcus Rashford and Anthony Martial before Fred's late clincher was as fine a tonic as the three points which keep them on the heels of Tottenham.	Football
Andy Murray and his brother Jamie beat Neal Skupski and Dan Evans in straight sets in Aberdeen but Scotland fall to 8-4 defeat against England; Andy Murray will head to the Australian Open encouraged by his performances. Andy and Jamie Murray were victorious in possibly their final match together in Scotland but England walked away with the inaugural Schroders Battle of the Brits trophy. Organised by Jamie Murray, the Scotland versus England clash proved a hit in Aberdeen, with around 20,000 fans attending the P&J Live arena across two days.	Tenis
The 6ft 1in right-handed Austrian has always had the game to do well on the surface with a booming backhand, stamina, graceful movement and aggressive play from the baseline. To prove a point, he conquered Nadal in the Barcelona Open semi-finals to become the first player other than Novak Djokovic to defeat the Spaniard on clay four times before capping an impressive week by crushing Daniil Medvedev 6-4 6-0 to win the title. He takes on 33-year-old Nadal in a repeat of last year's Roland Garros final but the Austrian will be feeling far more confident he can give the Spaniard a run for his money after ending Novak Djokovic's 26-match winning streak at Grand Slams. He is Austria's biggest tennis star since Thomas Muster, and now the 25-year-old is playing a free-flowing game under the tutelage of Olympic gold medallist Nicolas Massu.	Tenis
No one finished 2022 stronger than Novak Djokovic—he won 18 of his last 19 matches of the year, picking up titles at Tel Aviv, Astana and the ATP Finals (and reaching another final in Paris). He'll begin his 2023 season in Adelaide before heading to the Australian Open, and if his past results in Australia are anything to go by, get ready for that end-of-2022 momentum to continue. He's won his last 29 matches in a row in Australia. He's 27-5 against Top 10 players in Australia.	Tenis
Whether Tiger Woods makes it to Italy or not for next year's Ryder Cup, he will be an integral member of the U.S. team, captain Zach Johnson said Tuesday at the year-to-go ceremonies. Woods was a player on eight Ryder Cup teams and a vice-captain in 2018, before breaking bones in his right leg and ankle in a February 2021 car crash outside Los Angeles. The U.S. romped to a record 19-9 rout of Europe at Whistling Straits, Wisconsin, last year, and also won the Presidents Cup comfortably last month.	Golf
Adam Scott has confirmed he will join compatriots Cameron Smith and Marc Leishman at the Australian PGA Championship in November; LIV Golfers will be allowed to compete despite the event being co-sanctioned by the DP World Tour. Former Masters champion Adam Scott said he does not see the controversial LIV Series as 'pure evil' for the game of golf and called on the Saudi-backed circuit and the PGA Tour to move on from their feud. LIV Golf has lured away some of golf's biggest names with huge sums of money, while those who joined the rebel circuit were suspended by the PGA Tour. The breakaway circuit has filed a lawsuit accusing the PGA Tour of antitrust violations, while the PGA Tour Tour has filed a counterclaim.	Golf
The fact that we were able to do something under the lights that never before happened in our sport, to be able to grow it in a different way. Using just a 5-wood on a one-club challenge on the 455-yard, par-4 fourth hole, JT somehow makes par. JT and Spieth are 3-up after four holes.	Golf
The Spurs say they've sold more than 50,000 tickets for their Jan. 13 game against Golden State, which will be played in their former home venue, the Alamodome. A record 62,046 fans watched in 1998 when the Chicago Bulls visited the Atlanta Hawks.	Basketball
Coming off their NBA Finals run in 2021, the Suns were expected to potentially get back there again in 2022, but ultimately flamed out against Luka Doncic and the Dallas Mavericks in the second round. It was a painful loss for, not only us Suns fans but all NBA fans, as a Suns vs Warriors series in the Western Conference Finals would have been immense. Although the season did end on that sour note, it was in fact a historically great season for Phoenix. They were far and away the most consistent team all season boasting a top-five offensive and defensive rating. This efficiency helped them win a franchise-record 64 games, leading the NBA, with the next closest team being the Grizzlies with 56 wins.	Basketball
James Harden has vowed to get back to his high-scoring best for the Philadelphia 76ers after instructing president of basketball operations Daryl Morey to build a contender around him and Joel Embiid. The former league MVP said he would be happy to take a pay cut to allow the team to build a title-contending roster and he's been true to his word, agreeing a cut-price deal which has helped allow Morey time to put the pieces together to allow the Sixers to go full-tilt at winning a championship. Harden has now fully agreed to a two-year, $68.6m deal with the Philadelphia 76ers, ESPN reported on Wednesday. He averaged 22 points this past season for Brooklyn and Philadelphia, the lowest since he became a starter in the 2012-13 season. He turns 33 in August.	Basketball

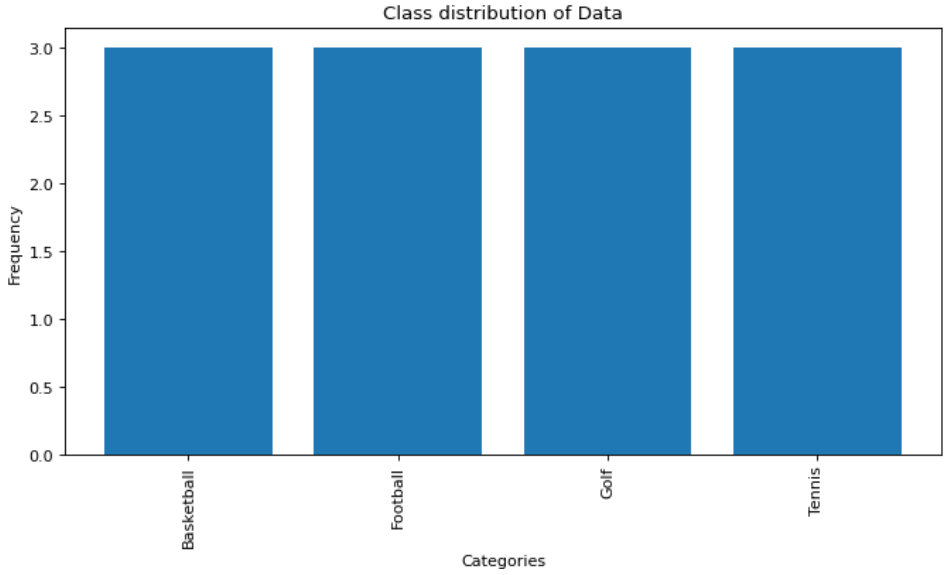
Hình 3 Các doc trong bộ dữ liệu

Dữ liệu được chia làm 2 phần:

- Train: 48 doc gồm 190 câu, 4834 từ được thu thập từ các trang web: Skysport, CNN, talksport.com .

Hình 4 Ngữ liệu huấn luyện (train)

- Test: 12 doc gồm 46 câu, 1259 từ được thu thập từ các trang web: goal.com, nba.com, eurosport.com, golfdigest.com .



Hình 5 Ngữ liệu kiểm tra (test)

2. Xử lý dữ liệu

2.1 Tách từ

Tách từ là bước tiền đề cho quá trình xử lý ngôn ngữ tự nhiên. Nó chia nhỏ câu thành từng từ, giúp cho việc phân tích hình thái từ trong câu dễ dàng hơn.

Tách từ đối với tiếng Anh khá đơn giản, ta chỉ việc sử dụng hàm split().

```
showz = 'Natural language processing strives to build machines that understand'
```

```
print('Before: ', showz)
print('After:  ', ', '.join(showz.split()))
```

```
Before:  Natural language processing strives to build machines that understand
After:   Natural, language, processing, strives, to, build, machines, that, understand
```

Hình 6 Tách từ tiếng Anh

2.2 Chuyển về từ gốc

Sau khi tách từ thì bước tiếp theo để phân tích hình thái từ là chuyển nó về từ gốc.

Trong tiếng Anh, các động từ và danh từ có thể ở nhiều thể, thì. Thế nên việc chuyển về từ gốc sẽ giúp kho từ của ta trở nên gọn gàng và tối ưu hơn, cùng với đó việc tính toán cũng trở nên nhanh hơn.

Ở bước này nhóm thực hiện 2 cách chuyển đổi từ về từ gốc: thủ công (không sử dụng thư viện hỗ trợ NLP) và sử dụng thư viện hỗ trợ NLP (thư viện NLTK).

❖ **Không sử dụng thư viện hỗ trợ NLP**

Việc chuyển đổi thủ công này nhóm làm thực hiện dựa trên bộ từ điển ***Cambridge*** [1]để giảm thiểu sai sót trong việc chuyển đổi bằng tay.

Hình 7 Chuyển về từ gốc thủ công

❖ **Sử dụng thư viện NLTK** [1]

Sử dụng hàm WordNetLemmatizer() với 4 nhãn

(pos='v', 'n', 'a', 'r')

```
lemmatizer = WordNetLemmatizer()
```

```
lemmaz = ' '.join([lemmatizer.lemmatize(word, pos='n') for word in showz.split()])
lemmaz = ' '.join([lemmatizer.lemmatize(word, pos='v') for word in lemmaz.split()])
lemmaz = ' '.join([lemmatizer.lemmatize(word, pos='a') for word in lemmaz.split()])
lemmaz = ' '.join([lemmatizer.lemmatize(word, pos='r') for word in lemmaz.split()])
```

Hình 8 Chuyển về từ gốc sử dụng thư viện NLTK

2.3 Loại bỏ từ phổ biến (stopwords)

Để tối ưu dữ liệu cho quá trình phân loại, ta có thể loại bỏ một số từ hay xuất hiện trong câu nhưng không quyết định đến việc phân loại

chủ đề. Ở bước này nhóm sử dụng stopword có sẵn của thư viện (stop_words='english').

3. Vector hóa văn bản

Máy tính không thể hiểu được ngôn ngữ bậc cao (tiếng Anh) nên ta cần phải chuyển đổi nó thành dạng ngôn ngữ mà máy tính có thể hiểu được. Và ngôn ngữ đó ở đây là dạng số. Các số sẽ biểu diễn tần suất xuất hiện của từ đó trong từng đoạn văn bản nhỏ.

Trong việc vector hóa văn bản này có một phương pháp khá dễ hiểu và dễ thực hiện đó là ***Bag of Word (BoW).***

> *Với BoW, nó là một phương pháp để trích xuất các đặc điểm từ các dữ liệu văn bản. Các đặc điểm này có thể được sử dụng để đào tạo các thuật toán học máy. 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. Hay nói cách khác, đó là một tập hợp bao gồm các cặp giá trị key và value, giá trị key là từ duy nhất có trong tập dữ liệu, giá trị value là số lần xuất hiện của từ đó trong câu, và BoW hầu như không quan tâm đến thứ tự xuất hiện của các từ đó.*

Hình 9 Thuật toán Bag of word

Biểu diễn trọng số của từ bằng tần suất xuất hiện khá đơn giản tuy nhiên nó không tối ưu, dễ bị gây nhiễu trong bài toán gồm 4 chủ đề cần phân loại đều nằm chung một lĩnh vực lớn. Vì vậy ta cần một kỹ thuật tính toán lại các con số trong ma trận tần suất để tối ưu, làm cho giá trị các số biểu diễn cho từng từ phản ánh đúng sự quan

trọng/không quan trọng của nó trong việc phân loại chủ đề. Và kỹ thuật đó là ***Term Frequency - Inverse Document Frequency (TF-IDF)***.

> *Ý tưởng của kỹ thuật: Đánh giá lại mức độ quan trọng trong tần suất xuất hiện của từ bằng cách:* ***số lần xuất hiện của từ đó trong một đoạn văn bản chia cho tổng số lần từ đó xuất hiện ở các văn bản có trong bộ ngữ liệu.***

Công thức:

$$\text{tfidf}(t,d,D) = \text{tf}(t,d) * \text{idf}(t,D) = \frac{f(t,d)}{\max\{f(w,d): w \in d\}} * \log\frac{|D|}{|\{d \in D: t \in d\}|}$$

Trong đó:

- tf(t, d): tần suất xuất hiện của từ t trong văn bản d.
- f(t, d): Số lần xuất hiện của từ t trong văn bản d
- max({f(w, d) : w ∈ d}): Số lần xuất hiện của từ có số lần xuất hiện nhiều nhất trong văn bản d
- idf(t, D): giá trị idf của từ t trong tập văn bản.
- |D|: Tổng số văn bản trong tập D
- |{d ∈ D : t ∈ d}|: thể hiện số văn bản trong tập D có chứa từ t.

4. Phân loại chủ đề sử dụng SVM model

Ở bước phân loại chủ đề, nhóm sử dụng mô hình SVM của thư viện scikit-learn để thực hiện huấn luyện và kiểm tra bộ ngữ liệu của mình.

Mô hình SVM [3] (Support Vector Machine) là một mô hình học có giám sát trong machine learning, được sử dụng để phân loại và dự đoán dữ liệu.

Ý tưởng cơ bản của mô hình SVM là tìm một siêu mặt phẳng (hyperplane) trong không gian nhiều chiều sao cho phân chia dữ liệu thành hai lớp sao cho khoảng cách giữa các điểm dữ liệu gần nhất đến siêu mặt phẳng đó là lớn nhất. Các điểm dữ liệu nằm gần nhất với siêu mặt phẳng được gọi là các vector hỗ trợ (support vectors), và chính vì

vậy, mô hình SVM còn được gọi là mô hình phân lớp bằng vector hỗ trợ.

sklearn.svm.SVC

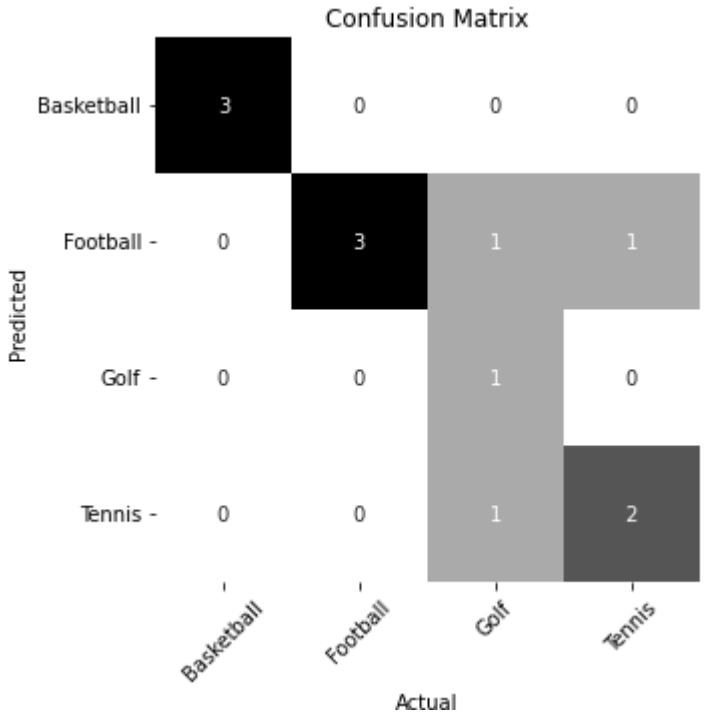
class **sklearn.svm.SVC**(*, C=1.0, kernel='rbf', degree=3, gamma='scale', coef0=0.0, shrinking=True, probability=False, tol=0.001, cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', break_ties=False,

Hình 10 Hàm SVC() với các tham số mặc định

III THỰC NGHIỆM VÀ ĐÁNH GIÁ

Bộ ngữ liệu dùng để huấn luyện và kiểm thử được nhóm thu thập từ hai nguồn báo điện tử riêng biệt để kết quả được công bằng hơn.

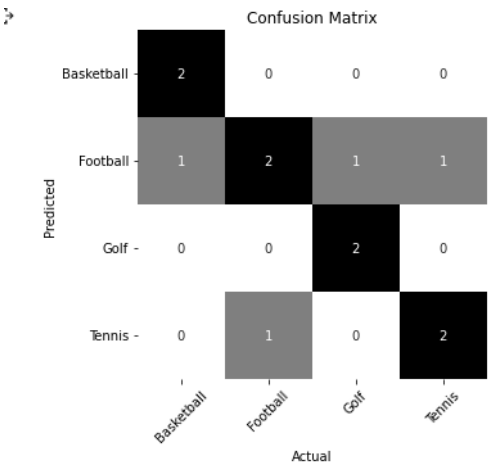
Với phương pháp phân loại chỉ sử dụng các phương pháp xử lý ngôn ngữ tự nhiên, không sử dụng 'stop words' thì kết quả thu được ở 2 cách chuyển từ gốc thủ công và sử dụng thư viện nltk đều cho accuracy = 0.75. Confusion Matrix của cả 2 cách chuyển từ gốc thủ công và sử dụng thư viện nltk cũng như nhau.



Hình 11 Confusion Matrix phương pháp phân loại có sử dụng NLP (không sử dụng stopwords)

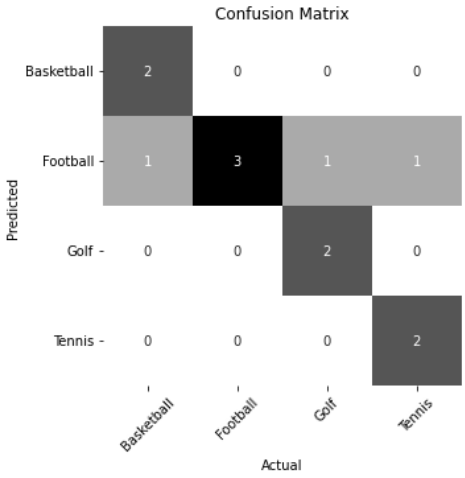
Với phương pháp phân loại chỉ sử dụng các phương pháp xử lý ngôn ngữ tự nhiên (sử dụng thêm stopwords của nltk) thì kết quả thu được ở 2 cách:

- Chuyển từ gốc thủ công: accuracy = 0.67
- Chuyển từ gốc sử dụng thư viện nltk: accuracy = 0.75.



Hình 12 Confusion Matrix

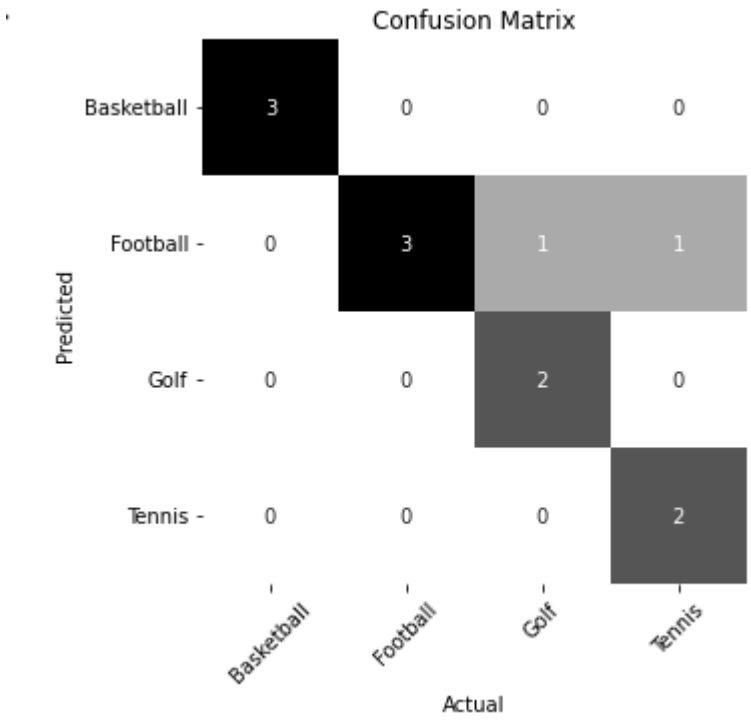
(Chuyển từ gốc thủ công)



Hình 13 Confusion Matrix

(Chuyển từ gốc với nltk)

Kết quả phân loại không sử dụng các phương pháp xử lý ngôn ngữ tự nhiên cho accuracy = 0.83, với Confusion matrix.



Hình 14 Confusion matrix (Không sử dụng các phương pháp xử lý ngôn ngữ tự nhiên)

Khi không sử dụng stopwords kết quả accuracy thu được sẽ **cao hơn** là có sử dụng.

Kết quả khi không sử dụng các phương pháp xử lý ngôn ngữ tự nhiên cho kết quả **cao nhất**.

Có 1 trường hợp dự đoán sai dù có sử dụng hay không sử dụng xử lý ngôn ngữ tự nhiên. Chủ đề **Tennis** nhưng mô hình dự đoán là **Football**.

Kyrgios sparked more controversy after pulling out of the United Cup tennis event at the eleventh hour, and drew a cheeky quip from Stefanos Tsitsipas. I know he doesn't like a lot to play Roland Garros. That's the only tournament that looks like he doesn't like to play that much. The rest of the tournaments, he played final in Wimbledon, and in New York he was close to the finals I think, with a very positive chance.

Hình 15 Trường hợp mô hình dự đoán là sai

Trong hình, từ '*United'* được xuất hiện nhiều trong chủ đề **Football** (các câu lạc bộ có United trong tên) ở dữ liệu huấn luyện dẫn đến việc mô hình dự đoán sai.

IV KẾT LUẬN

Kết quả phân loại giữa việc sử dụng dữ liệu được xử lý thủ công và thư viện là tương đương nhau, tuy nhiên việc xử lý dữ liệu bằng thư viện tiết kiệm thời gian hơn rất nhiều.

Bộ ngữ liệu còn ít để thấy được sự khác biệt rõ ràng về độ chính xác giữa việc xử lý ngữ liệu thủ công và xử lý ngữ liệu bằng thư viện.

➔ Độ chính xác phân loại chưa cao.

Hướng cải thiện:

- Bổ sung thêm dữ liệu.
- Tối ưu tập giá trị trong stopwords cho phù hợp với bài toán.

Tỉ lệ đóng góp (%)

	Lê Võ Tiến Phát	**Nguyễn Thành Trung**
Thu thập, xử lý dữ liệu	50	50
Thực nghiệm	50	50
Thiết kế power point	50	50
Viết report	50	50

Tài liệu tham khảo

[1] Cambridge Dictionary

[2] Thư viện nltk

[3] SVM Model

[4] Text classification using Python and scikit-learn