BỘ GIÁO DỤC VÀ ĐÀO TẠO

ĐẠI HỌC KINH TẾ TP HỒ CHÍ MINH

TRƯỜNG CÔNG NGHỆ VÀ THIẾT KẾ

ĐỒ ÁN MÔN HỌC

ĐỀ TÀI:

Phân tích tình cảm, trích xuất chủ đề và tóm tắt văn bản trong lĩnh vực Khoa học dữ liệu trên nền tảng Youtube

Học phần: DỮ LIỆU LỚN và ỨNG DỤNG

Nhóm Sinh Viên:

1. ĐỒNG ĐAN HOÀI
2. NGUYỄN QUỲNH KHÁNH HÀ
3. HUỲNH TRẦN ANH THY

Chuyên Ngành: KHOA HỌC DỮ LIỆU

Khóa: K46

Giảng Viên: ĐẶNG NHÂN CÁCH

TP. Hồ Chí Minh, Ngày 1 tháng 4 năm 2023

MỤC LỤC NỘI DUNG

MỤC LỤC HÌNH ẢNH

CHƯƠNG 1: GIỚI THIỆU

1.1. Giới thiệu bài toán và lý do chọn đề tài

Được thành lập vào năm 2005, YouTube đã phát triển trở thành công cụ tìm kiếm lớn thứ hai trên thế giới (sau Google), xử lý hơn 3 tỷ lượt tìm kiếm mỗi tháng [1]. Đối với những người sáng tạo nội dung trên YouTube, việc khai thác bình luận video để nhận biết được cảm xúc phản hồi của người xem về video là nguồn thông tin quan trọng, giúp chủ kênh có thể dễ dàng xác định được thái độ và sự hài lòng của người xem đối với nội dung và video truyền tải.

Những yếu tố dẫn đến sự thành công của một video trên Youtube như thời lượng video, độ dài tiêu đề, tần suất và ngày đăng tải video có tương quan thế nào đến các chỉ số thống kê về lượt thích, lượt bình luận, lượt view và tương tác của người xem cũng được các nhà sáng tạo nội dung quan tâm để xây dựng phát triển kênh YouTube của mình.

Ngoài ra cũng đáng để thử nghiệm và tìm kiếm xu hướng trong các chủ đề mà các kênh Youtube đang phủ sóng trong một chủ đề, lĩnh vực, thị trường ngách nhất định.

Là nhóm sinh viên chuyên ngành Khoa học dữ liệu, việc tìm hiểu các kênh Youtube sáng tạo nội dung về các chủ đề liên quan đến việc Phân tích dữ liệu và Khoa học dữ liệu là một điều cần thiết, để từ đó giúp sinh viên có cái nhìn bao quát về những xu hướng trong ngành hiện nay.

Vì vậy phạm vi của dự án phân tích Youtube này của nhóm sẽ chỉ tập trung vào phân tích các kênh Youtube liên quan đến lĩnh vực Khoa học dữ liệu mà sẽ không xem xét các lĩnh vực khác (vì mỗi lĩnh vực sẽ có những đặc thù về đối tượng người xem khác nhau). Do đó trong dự án này, nhóm sẽ khám phá dữ liệu thống kê của 7 channel YouTube là những kênh sáng tạo nội dung tiêu biểu về chủ đề Phân tích dữ liệu/Khoa học dữ liệu để từ đó tìm ra các hiểu biết sâu sắc về xu hướng ngành khoa học dữ liệu trên nền tảng YouTube, khai thác cảm xúc của người xem, và tóm tắt nội dung của video để từ đó giúp người xem dễ dàng xác định được nội dung liên quan đến chủ đề mà họ quan tâm.

1.2. Mục đích phân tích

Trong dự án này, nhóm sẽ thực hiện những điều sau:

- Thực hiện Data Crawling bằng YouTube API để lấy dữ liệu video của 7 channel YouTube làm nội dung về chủ đề *phân tích dữ liệu, khoa học dữ liệu*
- Phân tích dữ liệu video và xác định những "lầm tưởng" phổ biến về yếu tố giúp video hoạt động tốt trên YouTube, chẳng hạn như:
    + Số lượt thích và bình luận có quan trọng đối với một video để có được nhiều lượt xem hơn không?
    + Thời lượng video có quan trọng đối với lượt xem và tương tác (thích/ bình luận) không?
    + Độ dài tiêu đề có quan trọng đối với lượt xem không?
    + Đối với 7 kênh YouTube mà nhóm lựa chọn phân tích, tần suất họ tải video mới thường tập trung vào những ngày nào trong tuần?
- Phân tích cảm xúc những bình luận của người xem video bằng VADER, TextBlob và SentiWordNet. Gán nhãn bình luận tích cực/tiêu cực/trung lập (positive/negative/neutral). Đánh giá và so sánh 3 mô hình. Từ đó tìm ra channel nào nhận được nhiều bình luận tích cực nhất.
- Khám phá những chủ đề thịnh hành trong hơn 1000 video của 7 channel bằng các kỹ thuật xử lý ngôn ngữ tự nhiên và phương pháp phân cụm đối với tiêu đề video (title)
- Thực hiện tóm tắt văn bản đối với phần mô tả (description) của các video bằng phương pháp BERT.
- Tìm ra top 10 người thường xuất hiện nhất trong comment video

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

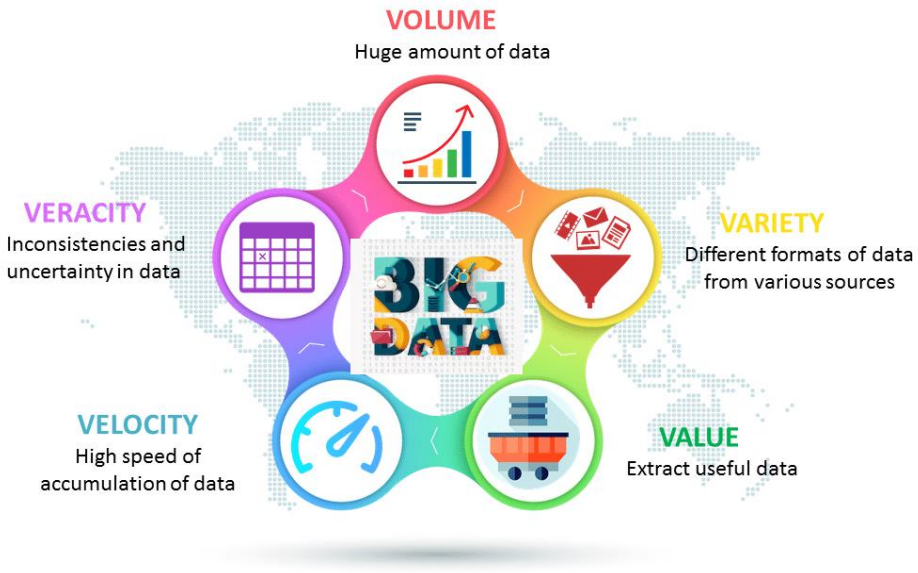
2.1. Big Data

a. Big Data là gì?

Big Data [2] là một thuật ngữ rộng cho việc xử lý một tập hợp dữ liệu rất lớn và phức tạp mà các ứng dụng xử lý dữ liệu truyền thống không xử lý được. Bao gồm các thách thức như phân tích, thu thập, giám sát dữ liệu, tìm kiếm, chia sẻ, lưu trữ, truyền nhận, trực quan hóa, truy vấn và tính riêng tư. Thuật ngữ Big Data thường được hiểu đơn giản là sử dụng để phân tích dự đoán hoặc là một số phương pháp tiên tiến khác rõ ràng để trích xuất giá trị từ dữ liệu mà ít khi đề cập đến kích thước của bộ dữ liệu. Độ chính xác trong Big Data có thể dẫn tới ra quyết định đúng đắn hơn, và những quyết định tốt hơn có thể đưa đến kết quả hoạt động tốt hơn như giảm chi phí và rủi ro.

b. Đặc tính

Big data có thể được mô tả bởi các đặc tính sau:



Hình 1. 5V - Bigdata [3]

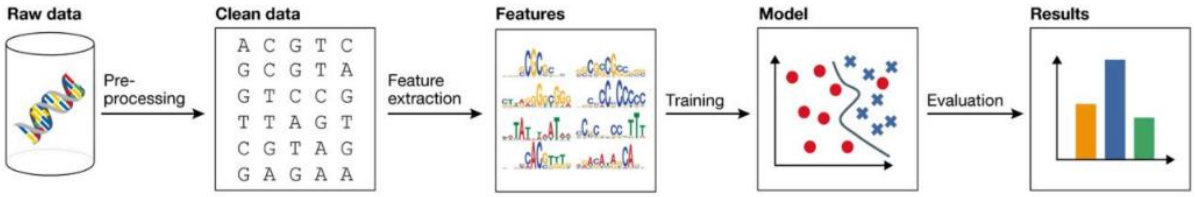
- *Volume (Tổng dung lượng lưu trữ):* Nói về số lượng dữ liệu được tạo và lưu trữ. Kích thước của dữ liệu sẽ được đánh giá là có giá trị và có tiềm năng hay không, và để xem xét liệu nó có thể được coi là dữ liệu lớn hay không.
- *Variety (Đa dạng kiểu dữ liệu):* Khái niệm này nói về type of data (kiểu dữ liệu) và nature of data (tính chất của dữ liệu). Điều này giúp những người

phân tích nó sử dụng hiệu quả thông tin chi tiết về kết quả. Chúng được tập hợp từ những text (văn bản), image (hình ảnh), sound (âm thanh), video; cộng với nó hoàn thành phần còn thiếu thông qua những thuật toán tổng hợp dữ liệu.

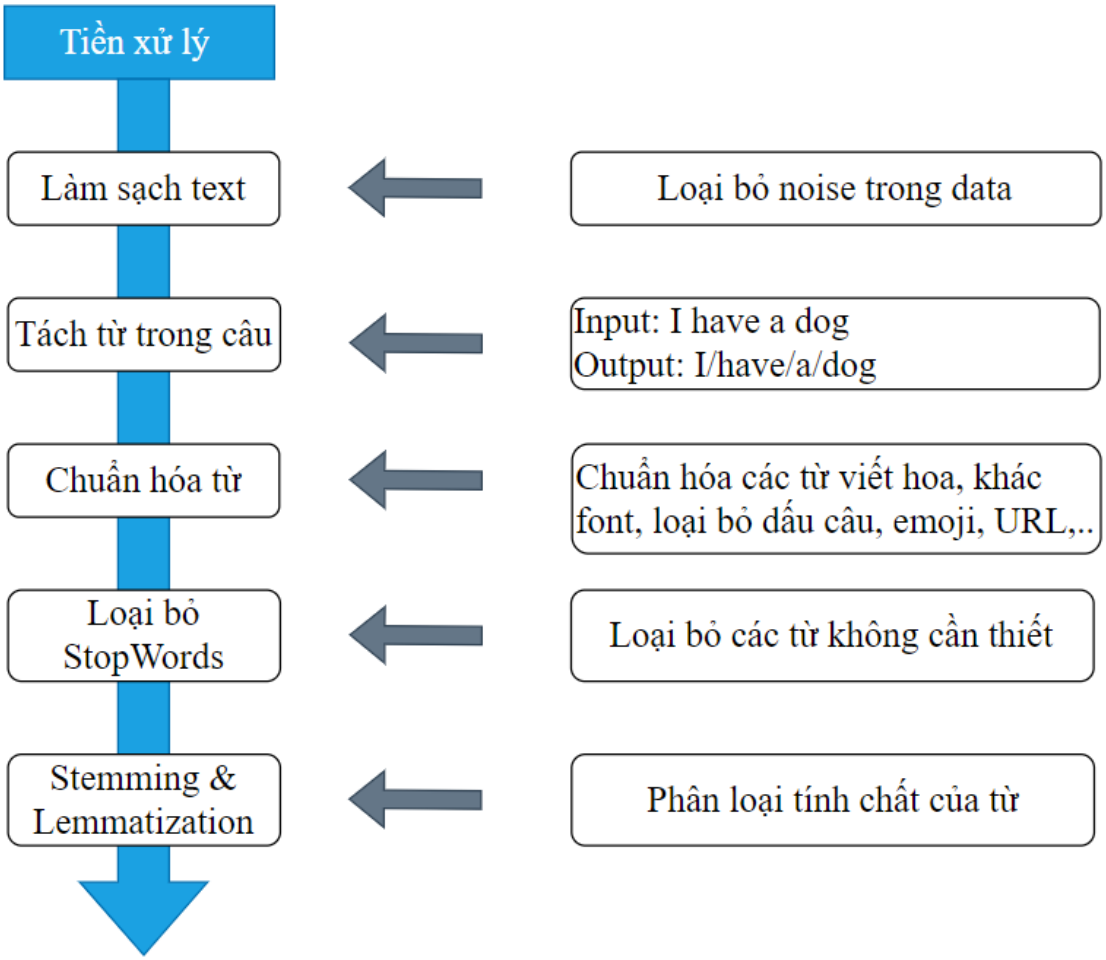
- *Value (Mức độ giá trị của thông tin):* Chất lượng dữ liệu của những dữ liệu lấy được có thể thay đổi rất nhiều, điều này sẽ ảnh hưởng rất mạnh đến việc phân tích chính xác những đấy. Ta có thể xem đây là tính chất cũng là khái niệm mà những doanh nghiệp hay nhà nghiên cứu muốn sử dụng và khai thác Big Data phải nắm giữ và am hiểu nó đầu tiên.
- *Velocity (Khả năng xử lý tốc độ cao):* Trong thời đại ngày nay, tốc độ dữ liệu được tạo ra và xử lý để đáp ứng nhu cầu và thách thức nằm trong con đường tăng trưởng và phát triển. Dữ liệu lớn thường có sẵn trong thời gian thực.
- *Veracity (Độ chính xác):* Vì đa dạng về các kiểu dữ liệu, nên sự không thống nhất của tập dữ liệu có thể cản trở các quy trình để xử lý và quản lý nó. Do đó, độ chính xác của công nghệ này có thể đảm bảo giúp cho việc giảm bớt sự sai lệch đáng tiếc có thể xảy ra.

2.2. Các phương pháp tiền xử lý

Bước đầu tiên và không thể thiếu trong việc xử lý ngôn ngữ tự nhiên là tiền xử lý. Vì văn bản vốn dĩ được liệt kê mà không có cấu trúc, để nguyên vậy để xử lý là rất khó khăn. Đặc biệt là loại văn bản trên web có lẫn các HTML tag, code JS, đó chính là noise.



Hình 2. Các giai đoạn tiền xử lý dữ liệu [4]



Hình 3. Các bước tiền xử lý NLP

Bước 1: Làm sạch text

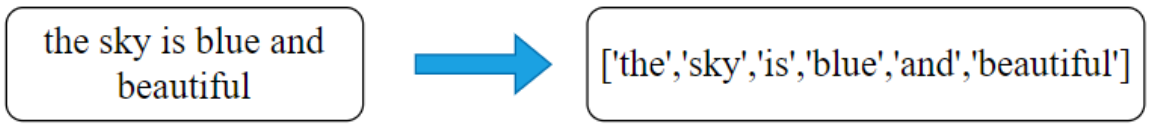
Mục đích bước này là loại bỏ noise trong data. Đầu tiên là biến đổi các câu thành kiểu chữ thường để thao tác loại bỏ các ký tự. Đa phần là loại bỏ các thẻ HTML và JS cũng có thể là những cụm từ không cần thiết, hay ký tự không có ý nghĩa ($%&##").

Hình 4. Làm sạch text

Bước 2: Tách từ trong câu

Để cấu thành nên đoạn văn bản bao giờ cũng bao gồm nhiều câu thành phần. Chính vì vậy để thực hiện các thao tác trên bộ dữ liệu text thì đầu tiên cần tách

riêng biệt các câu thành phần. Sử dụng hàm Tokenize để tách văn bản thành những câu riêng biệt. Thư viện NLTK cũng cung cấp hàm để tách các từ nltk.word_tokenize một cách chính xác.



Hình 5. Tách từ trong câu

Bước 3: Chuẩn hoá từ

Mục đích là đưa văn bản từ các dạng không đồng nhất về cùng một dạng. Dưới góc độ tối ưu bộ nhớ lưu trữ và tính chính xác cũng rất quan trọng.

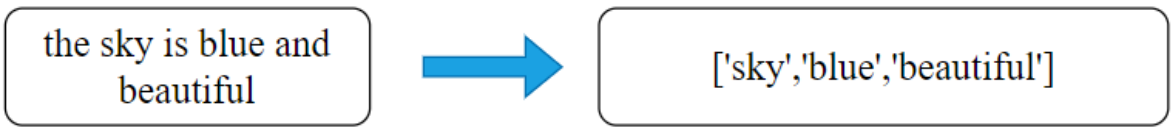
Ví dụ: U.S.A = USA

Thực hiện loại bỏ các ký tự đặc biệt như dấu câu như ?, !, ", ;,v.v, loại bỏ các emoji và URL (nếu có)

Bước 4: Loại bỏ StopWords

Stop words thường là các từ xuất hiện nhiều lần và không đóng góp nhiều vào ý nghĩa của câu, chúng sẽ đóng vai trò như nhiễu, trong tiếng Anh các từ này có thể kể đến như *the, is, at, on, which, in, some, many* hay trong tiếng Việt là các từ *cái, các, cả,....*

Để sử dụng stop words của NLTK, trước tiên ta cần download bộ stop words nltk.download('stopwords').



Hình 6. Loại bỏ StopWords

Bước 5: Stemming & Lemmatization

- **Stemming** là quá trình biến đổi các từ về dạng gốc của nó (ví dụ: connected, connection khi stemming thu được connect hay moved, move khi stemming thu được mov). Nhưng chưa chắc từ này có trong từ vựng

tiếng anh. Trong thư viện NLTK cũng có hỗ trợ thuật toán Porter để thực hiện nhiệm vụ này.

- **Lemmatization** về cơ bản là giống với stemming khi nó loại bỏ phần đuôi của từ để thu được gốc từ (ví dụ từ moved sau khi lemmatize sẽ thu được move). Chắc chắn từ này có trong từ vựng tiếng anh. Sử dụng part-of-speech tagging (nltk.pos_tag) để thu được các tính chất của từ.

2.3. Độ đo TF-IDF

TF-IDF [5] (Term Frequency – Inverse Document Frequency) là một kĩ thuật sử dụng trong khai phá dữ liệu văn bản. Trọng số này được sử dụng để đánh giá tầm quan trọng của một từ trong một văn bản. Giá trị cao thể hiện độ quan trọng cao và nó phụ thuộc vào số lần từ xuất hiện trong văn bản nhưng bù lại bởi tần suất của từ đó trong tập dữ liệu. TF-IDF cũng được sử dụng để lọc những từ stopwords trong các bài toán như tóm tắt văn bản và phân loại văn bản.

TF: *Term Frequency* (Tần suất xuất hiện của từ) là số lần từ xuất hiện trong văn bản. Vì các văn bản có thể có độ dài ngắn khác nhau nên một số từ có thể xuất hiện nhiều lần trong một văn bản dài hơn là một văn bản ngắn. Như vậy, term frequency thường được chia cho độ dài văn bản(tổng số từ trong một văn bản).

$$tf(t,d) = \frac{f(t,d)}{max\{f(w,d): w \epsilon d} \quad (1)$$

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: *Inverse Document Frequency* (Nghịch đảo tần suất của văn bản), giúp đánh giá tầm quan trọng của một từ . Khi tính toán TF , tất cả các từ được coi như có độ quan trọng bằng nhau. Nhưng một số từ như “is”, “of” và “that” thường

xuất hiện rất nhiều lần nhưng độ quan trọng là không cao. Như thế chúng ta cần giảm độ quan trọng của những từ này xuống.

$$idf(t, D) = log \frac{|D|}{|\{d \epsilon D: t \epsilon d\}|} \quad (2)$$

Trong đó:

- 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.

Cơ số logarit trong công thức này không thay đổi giá trị idf của từ mà chỉ thu hẹp khoảng giá trị của từ đó. Thay đổi cơ số sẽ không ảnh hưởng đến tỷ lệ giữa các giá trị IDF. Việc sử dụng logarit nhằm giúp giá trị tf-idf của một từ nhỏ hơn, do chúng ta có công thức tính tf-idf của một từ trong một văn bản là tích của tf và idf của từ đó.

Cụ thể, chúng ta có công thức tính tf-idf hoàn chỉnh như sau:

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D) \quad (3)$$

Những từ có giá trị TF-IDF cao là những từ xuất hiện nhiều trong văn bản này, và xuất hiện ít trong các văn bản khác. Việc này giúp lọc ra những từ phổ biến và giữ lại những từ có giá trị cao (từ khoá của văn bản đó).

2.4. Sentiment Analysis

Phân tích cảm xúc - Sentiment analysis [6] (được biết đến như là khai thác ý kiến - opinion mining hoặc trí tuệ nhân tạo cảm xúc - emotion AI) là việc sử dụng xử lý ngôn ngữ tự nhiên, phân tích văn bản, thuật ngữ máy tính, và sinh trắc học để xác định, truy xuất, định lượng, và nghiên cứu các trạng thái hiệu quả và thông tin chủ quan một cách có hệ thống. Phân tích cảm xúc được ứng dụng rộng rãi đến tài liệu tiếng nói của khách hàng như đánh giá, phản hồi khảo sát, phương tiện truyền thông xã hội và trực tuyến, tài liệu chăm sóc sức khỏe cho những ứng dụng phạm vi từ việc tiếp thị dịch vụ khách hàng đến y học lâm sàng. Với sự gia tăng của những mô hình học sâu, chẳng hạn như là RoBERTa, những miền dữ liệu khó hơn cũng có thể phân tích được, ví dụ những văn bản tin tức nơi mà các tác giả thường bài tỏ quan điểm/cảm xúc ít rõ ràng hơn.

2.5. VADER

a. Khái niệm

VADER [7] (Valence Aware Dictionary and sEntiment Reasoner) là một công cụ phân tích cảm xúc dựa trên quy tắc và từ vựng, đặc biệt phù hợp với cảm xúc được thể hiện trên mạng xã hội.

b. Mô tả nguồn tài liệu và bộ dữ liệu

Bao gồm TÀI NGUYÊN CHÍNH (mục 1-3) cũng như TÀI NGUYÊN THỬ NGHIỆM VÀ BỘ DỮ LIỆU bổ sung (mục 4-12)

i. vader_icwsm2014_final.pdf
ii. vader_lexicon.txt
iii. vaderSentiment.py
iv. tweets_GroundTruth.txt
v. tweets_anonDataRatings.txt
vi. nytEditorialSnippets_GroundTruth.txt
vii. nytEditorialSnippets_anonDataRatings.txt
viii. movieReviewSnippets_GroundTruth.txt
ix. movieReviewSnippets_anonDataRatings.txt
x. amazonReviewSnippets_GroundTruth.txt
xi. amazonReviewSnippets_anonDataRatings.txt
xii. Comp.Social website with more papers/research:

c. Việc chấm điểm

Điểm tổng hợp được tính bằng cách tính tổng các điểm hóa trị của mỗi từ trong từ vựng, được điều chỉnh theo các quy tắc và sau đó được chuẩn hóa thành giữa -1 (âm cực nhất) và +1 (cực dương nhất). Đây là số liệu hữu ích nhất nếu bạn muốn có một thước đo tình cảm đơn chiều cho một câu nhất định. Gọi nó là "điểm tổng hợp được chuẩn hóa, có trọng số" là chính xác.

- Tình cảm tích cực: điểm gộp >= 0,05
- Tình cảm trung lập: (điểm tổng hợp > -0,05) và (điểm tổng hợp < 0,05)
- Tình cảm tiêu cực: điểm tổng hợp <= -0,05

- **LƯU Ý**: Điểm tổng hợp là điểm được hầu hết các nhà nghiên cứu, bao gồm cả các tác giả, sử dụng phổ biến nhất để phân tích tình cảm.

	Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics			Ordinal Rank (by F1)		Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics		
		Overall Precision	Overall Recall	Overall F1 score				Overall Precision	Overall Recall	Overall F1 score
Social Media Text (4,200 Tweets)							**Movie Reviews (10,605 review snippets)**			
Ind. Humans	**0.888**	0.95	0.76	0.84	2	1	0.899	0.95	0.90	0.92
VADER	0.881	**0.99**	**0.94**	**0.96**	1*	2	0.451	0.70	0.55	0.61
Hu-Liu04	0.756	0.94	0.66	0.77	3	3	0.416	0.66	0.56	0.59
SCN	0.568	0.81	0.75	0.75	4	7	0.210	0.60	0.53	0.44
GI	0.580	0.84	0.58	0.69	5	5	0.343	0.66	0.50	0.55
SWN	0.488	0.75	0.62	0.67	6	4	0.251	0.60	0.55	0.57
LIWC	0.622	0.94	0.48	0.63	7	9	0.152	0.61	0.22	0.31
ANEW	0.492	0.83	0.48	0.60	8	8	0.156	0.57	0.36	0.40
WSD	0.438	0.70	0.49	0.56	9	6	0.349	0.58	0.50	0.52
Amazon.com Product Reviews (3,708 review snippets)							**NY Times Editorials (5,190 article snippets)**			
Ind. Humans	0.911	0.94	0.80	0.85	1	1	0.745	0.87	0.55	0.65
VADER	0.565	0.78	0.55	0.63	2	2	0.492	0.69	0.49	0.55
Hu-Liu04	0.571	0.74	0.56	0.62	3	3	0.487	0.70	0.45	0.52
SCN	0.316	0.64	0.60	0.51	7	7	0.252	0.62	0.47	0.38
GI	0.385	0.67	0.49	0.55	5	5	0.362	0.65	0.44	0.49
SWN	0.325	0.61	0.54	0.57	4	4	0.262	0.57	0.49	0.52
LIWC	0.313	0.73	0.29	0.36	9	9	0.220	0.66	0.17	0.21
ANEW	0.257	0.69	0.33	0.39	8	8	0.202	0.59	0.32	0.35
WSD	0.324	0.60	0.51	0.55	6	6	0.218	0.55	0.45	0.47

Table 4: VADER 3-class classification performance as compared to individual human raters and 7 established lexicon baselines across four distinct domain contexts (clockwise from upper left: tweets, movie reviews, product reviews, opinion news articles).

Hình 7. Chỉ số đánh giá của Vader [8]

Nhìn từ ảnh trên có thể thấy rằng, VADER có chỉ số đánh giá mức độ chính xác rất tốt khi so với của Ind. Humans (do từng cá nhân con người phân loại) ở 4 thể loại văn bản khác nhau, bao gồm: Social Media Text, Movie Reviews, Amazon.com Product Reviews, và NY Times Editorials. Khi 3 chỉ số Precision, Recall, F1-score của VADER đều lớn hơn của Ind. Humans ở văn bản Social Media. Ngoài ra, khi mà chỉ đứng vị trí thứ 2 (sau Ind. Humans) so với 7 công cụ khác ở 3 loại văn bản còn lại.

Có thể nói rằng VADER rất tốt để phân loại phân loại văn bản social media khi mà khi chỉ số Precision, Recall, và F1 lần lượt là 0.99, 0.94, và 0.96. Chúng cao vượt trội hơn so với con người và những công cụ khác phân loại.

2.6. SentiWordNet

SENTIWORDNET [9] là kết quả của việc chú thích tự động tất cả các tập hợp của WORDNET theo các khái niệm "tính tích cực", "tính tiêu cực" và "tính trung lập". Mỗi tập hợp từ đồng nghĩa được liên kết với ba điểm số Pos(s), Neg(s) và Obj(s) cho biết mức độ tích cực, tiêu cực và "khách quan" (nghĩa là trung tính) của các thuật ngữ chứa trong tập hợp đồng nghĩa.

	Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics			Ordinal Rank (by F1)		Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics		
		Overall Precision	Overall Recall	Overall F1 score				Overall Precision	Overall Recall	Overall F1 score
Social Media Text (4,200 Tweets)							**Movie Reviews (10,605 review snippets)**			
Ind. Humans	**0.888**	0.95	0.76	0.84	2	1	0.899	0.95	0.90	0.92
VADER	0.881	**0.99**	**0.94**	**0.96**	1*	2	0.451	0.70	0.55	0.61
Hu-Liu04	0.756	0.94	0.66	0.77	3	3	0.416	0.66	0.56	0.59
SCN	0.568	0.81	0.75	0.75	4	7	0.210	0.60	0.53	0.44
GI	0.580	0.84	0.58	0.69	5	5	0.343	0.66	0.50	0.55
SWN	0.488	0.75	0.62	0.67	6	4	0.251	0.60	0.55	0.57
LIWC	0.622	0.94	0.48	0.63	7	9	0.152	0.61	0.22	0.31
ANEW	0.492	0.83	0.48	0.60	8	8	0.156	0.57	0.36	0.40
WSD	0.438	0.70	0.49	0.56	9	6	0.349	0.58	0.50	0.52
Amazon.com Product Reviews (3,708 review snippets)							**NY Times Editorials (5,190 article snippets)**			
Ind. Humans	0.911	0.94	0.80	0.85	1	1	0.745	0.87	0.55	0.65
VADER	0.565	0.78	0.55	0.63	2	2	0.492	0.69	0.49	0.55
Hu-Liu04	0.571	0.74	0.56	0.62	3	3	0.487	0.70	0.45	0.52
SCN	0.316	0.64	0.60	0.51	7	7	0.252	0.62	0.47	0.38
GI	0.385	0.67	0.49	0.55	5	5	0.362	0.65	0.44	0.49
SWN	0.325	0.61	0.54	0.57	4	4	0.262	0.57	0.49	0.52
LIWC	0.313	0.73	0.29	0.36	9	9	0.220	0.66	0.17	0.21
ANEW	0.257	0.69	0.33	0.39	8	8	0.202	0.59	0.32	0.35
WSD	0.324	0.60	0.51	0.55	6	6	0.218	0.55	0.45	0.47

Table 4: VADER 3-class classification performance as compared to individual human raters and 7 established lexicon baselines across four distinct domain contexts (clockwise from upper left: tweets, movie reviews, product reviews, opinion news articles).

Nhìn từ ảnh trên có thể thấy rằng, Sentiwordnet (SWN) có chỉ số đánh giá mức độ chính xác tương đối trung bình khi so với của Ind. Humans(do từng cá nhân con người phân loại) ở 4 thể loại văn bản khác nhau, bao gồm: Social Media Text, Movie Reviews, Amazon.com Product Reviews, và NY Times Editorials. Khi 3 chỉ số Precision, Recall, F1-score của SWN lần lượt là 0.75, 0.62, 0.67 ở văn bản Social Media.

Có thể nói rằng SWN chưa đủ tin cậy để phân loại văn bản social media khi mà khi chỉ số Precision, Recall, và F1 còn thấp và đúng ở vị trí thứ 6 trong 9 công cụ trên.

2.6. TextBlob

Một số điều cơ bản về TextBlob [10], Nó sử dụng NLTK (Bộ công cụ ngôn ngữ tự nhiên) và đầu vào chứa một câu duy nhất, Đầu ra của TextBlob là tính phân cực và tính chủ quan. Điểm phân cực nằm trong khoảng từ (-1 đến 1) trong đó -1 xác định những từ tiêu cực nhất như 'ghê tởm', 'khủng khiếp', 'thảm hại' và 1 xác định những từ tích cực nhất như 'xuất sắc', 'tốt nhất'. Điểm chủ quan nằm trong khoảng từ (0 đến 1), Nó cho biết mức độ quan điểm cá nhân, Nếu một câu có tính chủ quan cao, tức là gần bằng 1, thì có vẻ như văn bản chứa nhiều quan điểm cá nhân hơn là thông tin thực tế.

Kết quả đạt được sau khi phân tích bộ dữ liệu có 1.4 triệu dữ liệu Tweets đã dán nhãn.

Algorithm	Accuracy
Textblob	56%
VADER	56%
Flair	50%
USE model	0.775

Hình 8. Bảng so sánh các chỉ số đánh giá

2.7. NER (Named entity recognition)

Nhận dạng thực thể được đặt tên (NER) [11] (còn được gọi là nhận dạng thực thể (được đặt tên), phân đoạn thực thể và trích xuất thực thể) là một nhiệm vụ con của trích xuất thông tin nhằm tìm cách định vị và phân loại các thực thể được đặt tên được đề cập trong văn bản phi cấu trúc thành các danh mục được xác định trước, chẳng hạn như người tên, tổ chức, địa điểm, mã y tế, biểu thức thời gian, số lượng, giá trị tiền tệ, tỷ lệ phần trăm, v.v.

2.8. K-Means

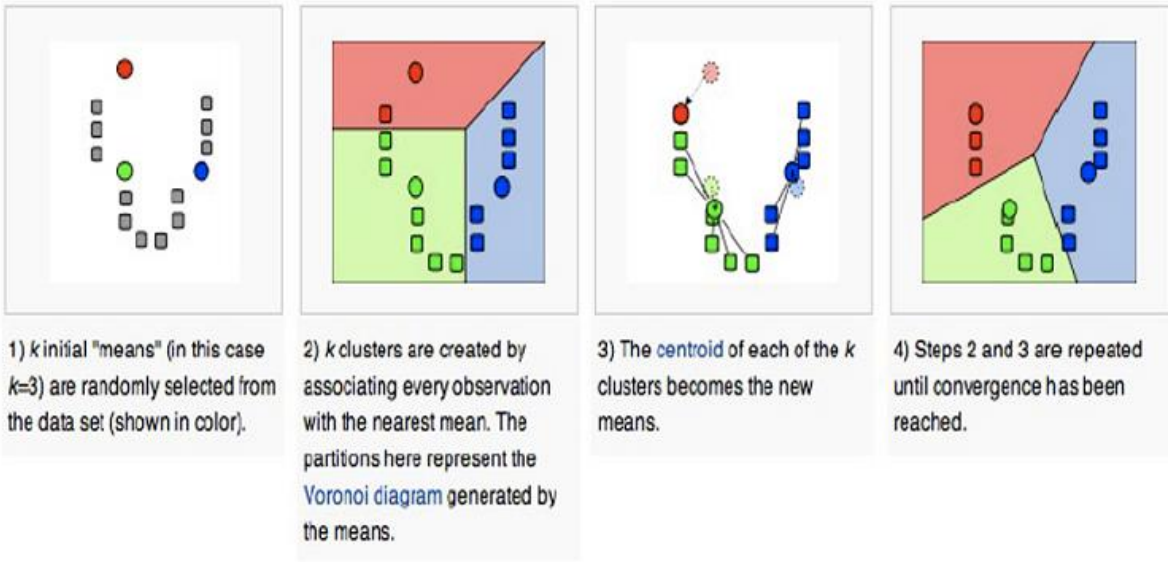
Trong học máy không giám sát (unsupervised learning), tập dữ liệu của chúng ta không được gán nhãn cụ thể. Phân cụm là một hình thức học không giám sát giúp đánh giá và gom các cá thể có chung kiểu mẫu (pattern) về một cụm. Kết quả cuối cùng thu được sẽ là những cụm cá thể, cá thể ở các cụm khác nhau sẽ khác biệt với nhau, các cá thể trong cùng một cụm sẽ tương đồng với nhau trên tiêu chuẩn đã được đặt ra.

K-Means là một phương pháp phân cụm cứng (hard clustering), trong đó mỗi cá thể sẽ chỉ thuộc về một cụm duy nhất hoặc không thuộc cụm nào. Trong đó ta phải lựa chọn trước số cụm trước khi tiến hành phân cụm. Mỗi lần phân cụm theo phần (partitional) chúng ta sẽ thu được các kết quả phân cụm khác nhau.

Trong phân cụm K-Means [12], ta sẽ có các bước cơ bản như sau:

1. Chọn ngẫu nhiên k trọng tâm (centroid) của các cụm

2. Dựa trên khoảng cách với các trọng tâm (centroid) của các cụm để phân phối và sắp xếp các cá thể (instance) vào các cụm.
3. Tính trọng tâm (centroid) mới của các cụm sau khi được sắp xếp
4. Lặp lại bước số 2 cho đến khi trọng tâm (centroid) của các cụm không thay đổi nữa

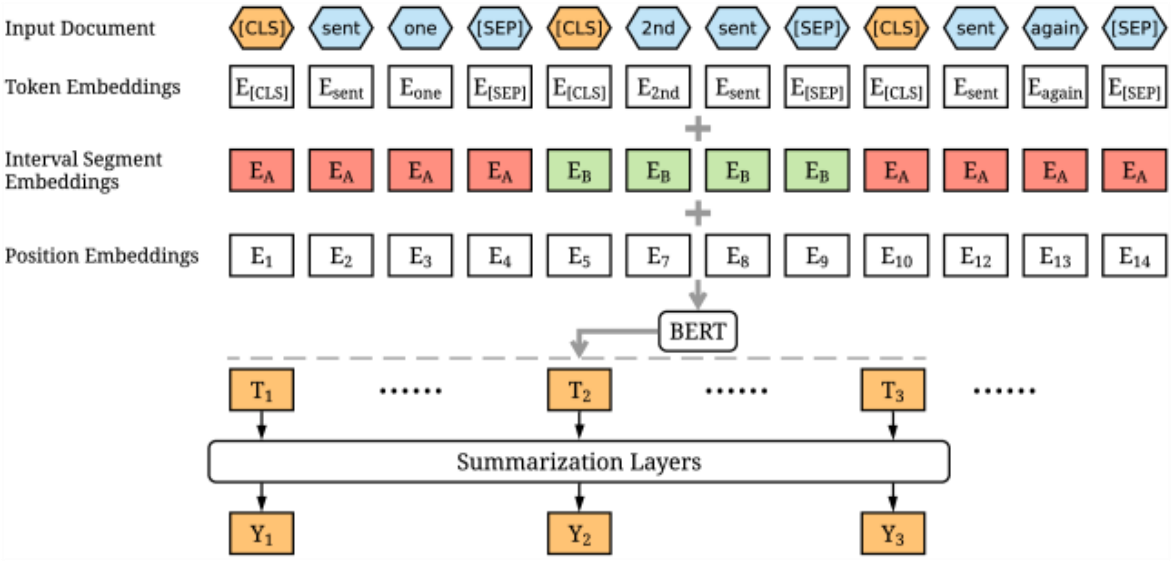


Hình 9. K-Means clustering [13]

Tiêu chuẩn để phân cụm của K-Means là khoảng cách của các điểm dữ liệu so với trọng tâm (centroid). Để tính khoảng cách này chúng ta có thể sử dụng các khoảng cách Euclidean, Weighted Euclidean/, Minkowski, Manhattan,... Phổ biến nhất vẫn là Euclidean.

2.9. BERT Extractive Summarizer

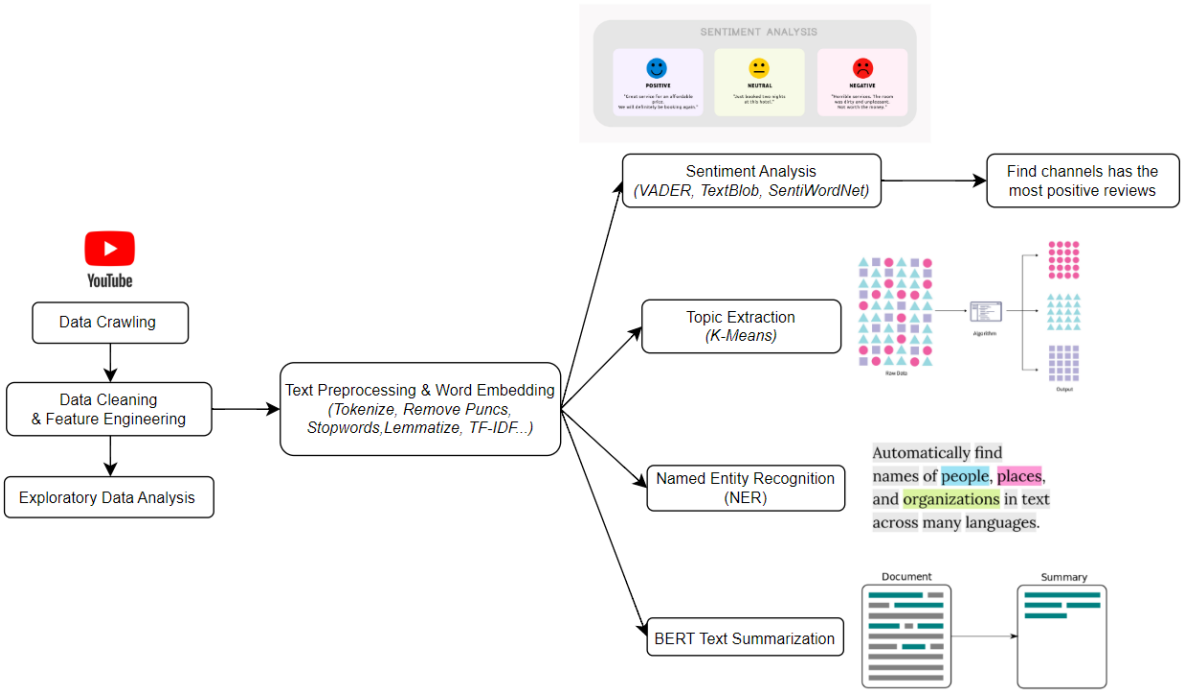
BERT [14] (Bidirectional Encoder Representations from Transformers) là một mô hình học máy sẵn hay còn gọi là pre-trained model, học ra các vector đại diện theo ngữ cảnh 2 chiều của từ, được sử dụng để transfer sang các bài toán khác trong lĩnh vực xử lý ngôn ngữ tự nhiên. BERT đã thành công trong việc cải thiện những công việc gần đây trong việc tìm ra đại diện của từ trong không gian số (không gian mà máy tính có thể hiểu được) thông qua ngữ cảnh của nó.
Từ đó được áp dụng trong bài toán tóm tắt rút trích văn bản. Các câu sẽ được biểu diễn dưới dạng vector đặc trưng sử dụng BERT, sau đó được phân lớp để chọn ra những câu quan trọng làm bản tóm tắt.



Hình 10. Tổng quan về mô hình BERTSUM [15]

CHƯƠNG 3: PHƯƠNG PHÁP THỰC HIỆN & DỮ LIỆU

3.1. Phương pháp thực hiện



Hình 11. Flow Chart

_ Đầu tiên nhóm thực hiện crawl data từ YouTube thông qua YouTube API. Với 7 kênh YouTube ta kết quả thu được dữ liệu statistics của 1079 videos và 19646 dòng comments.

_ Sau đó thực hiện làm sạch và tiền xử lý dữ liệu video statistics (Data Cleaning & Feature Engineering) và thực hiện phân tích khám phá dữ liệu (Exploratory Data Analysis). Tìm ra các yếu tố

_ Sử dụng các phương pháp tiền xử lý làm sạch dữ liệu văn bản (Text Preprocessing) để xử lý các biến 'comment', 'title', 'description'. Vector hóa văn bản bằng phương pháp TF-IDF (Word Embedding)

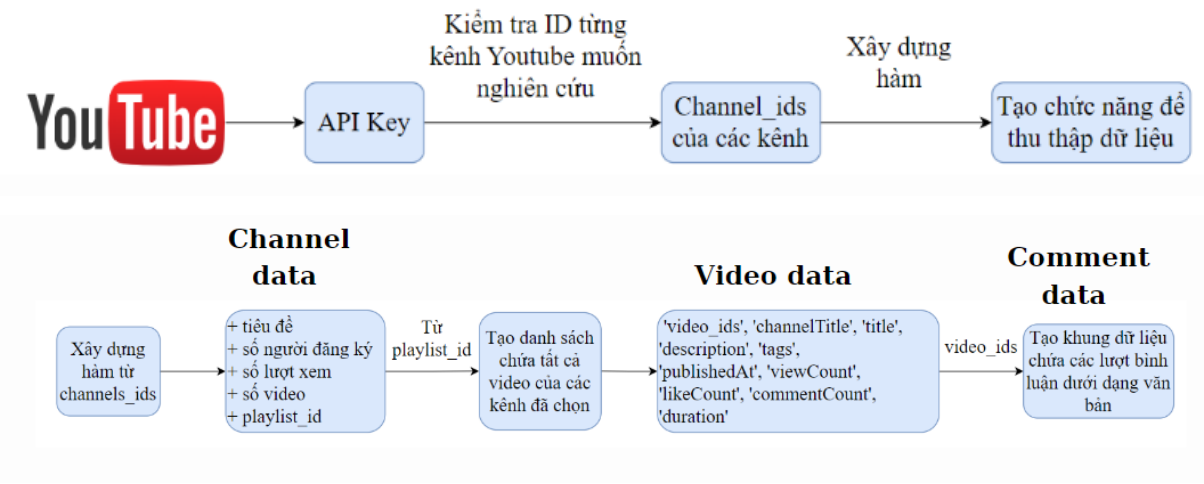
Từ vector văn bản, thực hiện các bước sau:

+ Phân loại cảm xúc comment (Sentiment Analysis) bằng 3 phương pháp lexicon-based VADER, TextBlob, SentiWordNet. Gán nhãn các comments theo positive/ negative/ neutral

  Tìm ra channel có nhiều lượng bình luận tích cực nhất dựa trên 3 phương pháp. So sánh sự khác nhau giữa 3 mô hình.
+ Thực hiện trích xuất các chủ đề của các video YouTube dựa trên biến 'title' bằng phương pháp phân cụm K-Means. Tìm ra các chủ đề nổi bật trong 1079 videos
+ Trích xuất top 10 người có tần suất xuất hiện cao nhất trong 'comment' bằng phương pháp Named Entity Recognition (NER).

+ Áp dụng BERT Extractive Summarizer để tóm tắt văn bản cho biến ‘description’ mô tả video

3.2. Data Crawling



Hình 12. Các bước Crawling data

Để có được bộ dữ liệu này, đầu tiên đăng nhập kênh Youtube cá nhân sau đó tiến hành lấy được API Key để thực hiện quá trình crawling data. Sau đó truy cập Youtube và kiểm tra ID channel của từng kênh muốn đưa vào phạm vi nghiên cứu (sử dụng URL của chúng). Sau đó, nhóm đã tạo các chức năng để lấy số liệu thống kê kênh thông qua API. Cụ thể lấy được những số liệu thống kê như sau:

- Xây dựng hàm thu thập thống kê về các kênh được chọn bao gồm các thông tin tiêu đề, số người đăng ký, số lượt xem, số video, playlist ID. Hàm được dựa trên thông số channels_ids: danh sách ID kênh.

→ Kết quả thu được:

Khung dữ liệu chứa thống kê kênh cho tất cả các kênh trong danh sách được cung cấp: tiêu đề, số người đăng ký, số lượt xem, số video, playlist ID.

```
channel_data = get_channel_stats(youtube, channel_ids)
channel_data
```

	channelName	subscribers	views	totalVideos	playlistId
0	Luke Barousse	321000	14528430	123	UULLw7jmFsvfIVaUFsLs8mlQ
1	365 Data Science	276000	11899300	215	UUEBpSZhI1X8WaP-kY_2LLcg
2	Tina Huang	464000	19524707	118	UU2UXDak6o7rBm23k3Vv5dww
3	Ken Jee	237000	7661622	263	UUiT9RITQ9PW6BhXK0y2jaeg
4	Alex The Analyst	399000	15474861	196	UU7cs8q-gJRlGwj4A8OmCmXg
5	The Almost Astrophysicist	22800	1095407	80	UUtC_WTVuo9k3ZoI0ZB6u5mQ
6	techTFQ	173000	8287427	84	UUnz-ZXXER4jOvuED5trXfEA

- Tiếp theo, tạo ra một danh sách chứa các ID video của tất cả video trong danh sách kênh đã cho. Thông số nhận được là playlist_id: ID video danh sách phát của kênh.

→ ***Kết quả thu được:***

Danh sách ID video của tất cả video trong danh sách phát của các kênh.

- Sau đó, xây dựng hàm dựa trên playlist_id để nhận số liệu thống kê video của tất cả các video có ID như trong danh sách.

→ ***Kết quả thu được:***

Khung dữ liệu với số liệu thống kê về video gồm: ‘video_ids’, ‘channelTitle’, ‘title’, ‘description’, ‘tags’, ‘publishedAt’, ‘viewCount’, ‘likeCount’, ‘commentCount’, ‘duration’.

```
video_df
```

	video_id	channelTitle	title	description	tags	publishedAt	viewCount	likeCount	commentCount	duration
0	jdqWbzwm1IU	Luke Barousse	THIS is what Hiring Managers look for - Pt.2	Full Video Here https://youtu.be/clZWgPmpRV...	[data viz by luke, business intelligence, data...	2023-03-03T16:00:32Z	1064	53	0	PT38S
1	BL1w5chqt8U	Luke Barousse	THIS is what Hiring Managers look for	Full Video Here https://youtu.be/clZWgPmpRV...	[data viz by luke, business intelligence, data...	2023-03-01T16:00:06Z	2598	198	4	PT39S
2	CTLgC4AaOtM	Luke Barousse	SQL... but for non-data nerds	Full Video Here youtu.be/GEBzsz8ZSXs\n\nCou...	[data viz by luke, business intelligence, data...	2023-02-27T16:00:14Z	8440	620	6	PT46S
3	Zgx6dwZRov4	Luke Barousse	Popular SQL databases for data nerds	Full Video Here youtu.be/GEBzsz8ZSXs\n\nCou...	[data viz by luke, business intelligence, data...	2023-02-24T16:00:17Z	7087	520	4	PT37S
4	tJAzELZv4gE	Luke Barousse	What is SQL?!?	Full Video Here youtu.be/GEBzsz8ZSXs\n\nCou...	[data viz by luke, business intelligence, data...	2023-02-22T16:00:37Z	9444	572	12	PT38S
...	...	...	...	...	...	...	...	...	...	...
1074	1aybOgni7II	techTFQ	How to install PostgreSQL on Mac OS | Install ...	This video is about how to install PostgreSQL ...	[PostgreSQL, pgAdmin, Database, PostgreSQL Dat...	2020-11-16T02:28:09Z	50410	519	90	PT20M51S
1075	j09EQ-xlh88	techTFQ	Learn What is Database | Types of Database | DBMS	In this video, we learn everything we need to ...	[Database, DBMS, Relational Database, Non-Rela...	2020-08-30T00:38:24Z	143605	2708	105	PT12M11S
1076	7nzTDrio7vY	techTFQ	Do you need a Smartwatch	In this video, I talk about the advantageous o...	[Do you need a smartwatch, Samsung Galaxy Watc...	2020-07-12T15:32:36Z	11125	182	46	PT7M43S
1077	J-uCLHTlWZ4	techTFQ	MacBook Pro 13 2020 One Week Later Review	Macbook Pro 13 2020 review after one week. Qui...	None	2020-06-29T15:03:19Z	1316	70	20	PT9M24S
1078	_BMPh5M4BIY	techTFQ	MacBook Pro 13 2020 Unboxing	This video is about unboxing of New MacBook Pr...	[#MacBookPro13, #Unboxing, #Apple, #Malaysia, ...	2020-06-22T14:49:50Z	2275	86	29	PT9M19S

1079 rows × 11 columns

- Cuối cùng là tạo hàm thu về các lượt bình luận dưới dạng văn bản từ tất cả các video có ID như trong danh sách thông qua thông số video_ids: danh sách ID video.

→ ***Kết quả thu được:***

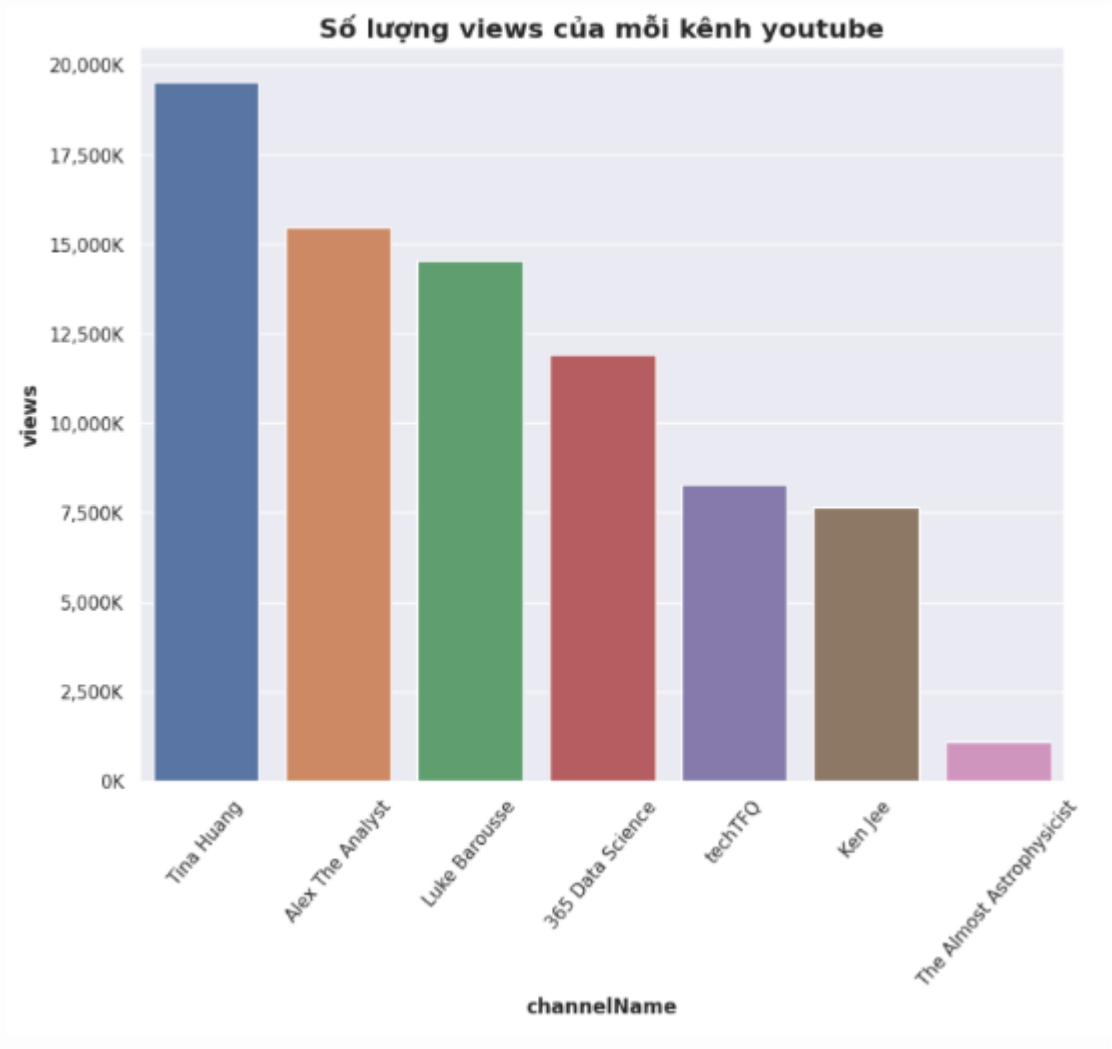
Khung dữ liệu có ID video và các lượt bình luận được thể hiện dưới dạng văn bản với biến ‘textOriginal’

```
comments_df = comments_df.drop_duplicates(subset=['textOriginal', 'commentId']).reset_index(drop=True)
comments_df
```

	videoId	textDisplay	textOriginal	parentId	authorDisplayName	authorProfileImageUrl
0	BL1w5chqt8U	Full video here: <a href="https://youtu.be/clZ...	Full video here: https://youtu.be/clZWgPmpRVc	UgycHUYXBSYRyfqOohR4AaABAg	Luke Barousse	https://yt3.ggpht.com/ytc/AL5GRJVV30j1gbwVosUG...
1	BL1w5chqt8U	Damn I hate shorts. No time no elaborate any i...	Damn I hate shorts. No time no elaborate any i...	None	icl	https://yt3.ggpht.com/ytc/AL5GRJVclyMZr7H59uuO...
2	BL1w5chqt8U	That's one checkbox, now the rest of the 9...	That's one checkbox, now the rest of the 99+ f...	None	Dark GT	https://yt3.ggpht.com/tpzGM83NYCSxEL9yOqHsZpPy...
3	CTLgC4AaOtM	Where does Microsoft Access in all this?	Where does Microsoft Access in all this?	None	Ty Alva	https://yt3.ggpht.com/r9hQCiyRUf6XmF_0J8dPBUXH...
4	CTLgC4AaOtM	check out my "How I use Excel" video...	check out my "How I use Excel" video; I go ove...	UgyxV5DG_PQTskCSd_t4AaABAg	Luke Barousse	https://yt3.ggpht.com/ytc/AL5GRJVV30j1gbwVosUG...
...	...	...	...	...	...	...
19641	_BMPh5M4BIY	Thank you Ashay	Thank you Ashay	Ugy9bvTqkwfBICIDwZJ4AaABAg	techTFQ	https://yt3.ggpht.com/68QpkOCRQespfOQ5yZwhCrM6...
19642	_BMPh5M4BIY	Thank you Shariq	Thank you Shariq	UgxhRMlzw8mZ_TQ6qR14AaABAg	techTFQ	https://yt3.ggpht.com/68QpkOCRQespfOQ5yZwhCrM6...
19643	_BMPh5M4BIY	Thank you Shahil	Thank you Shahil	UgznoxSfu68-tyPwuRF4AaABAg	techTFQ	https://yt3.ggpht.com/68QpkOCRQespfOQ5yZwhCrM6...
19644	_BMPh5M4BIY	Thank you Munawar	Thank you Munawar	UgwwfaLN-5Y4NxFPLcp4AaABAg	techTFQ	https://yt3.ggpht.com/68QpkOCRQespfOQ5yZwhCrM6...
19645	_BMPh5M4BIY	Thank you Sheeba	Thank you Sheeba	UgxDJXNQNHjaQ4ifSGt4AaABAg	techTFQ	https://yt3.ggpht.com/68QpkOCRQespfOQ5yZwhCrM6...

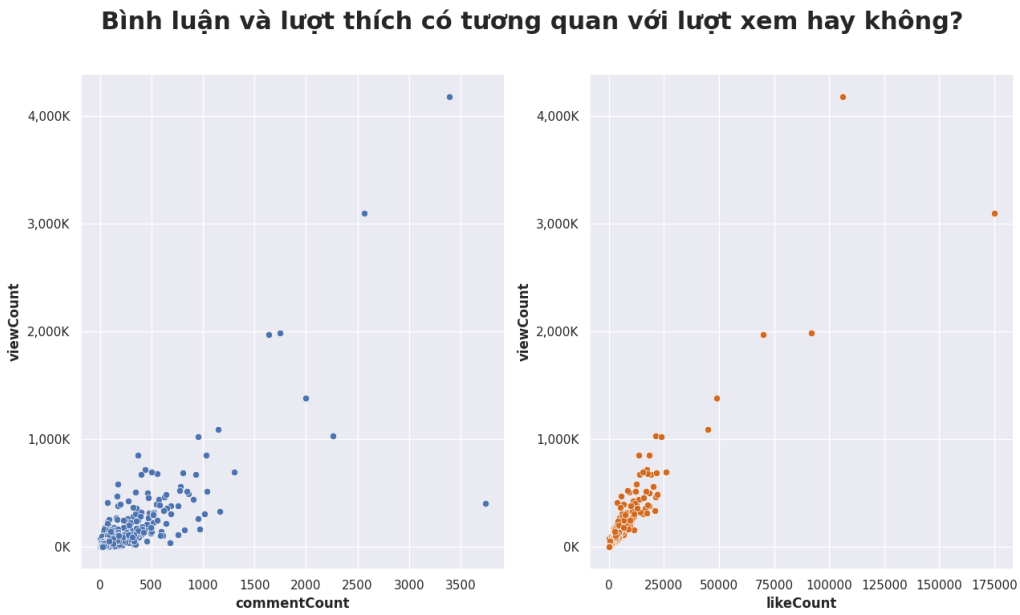
19646 rows × 15 columns

3.3. Exploratory Data Analysis (EDA)

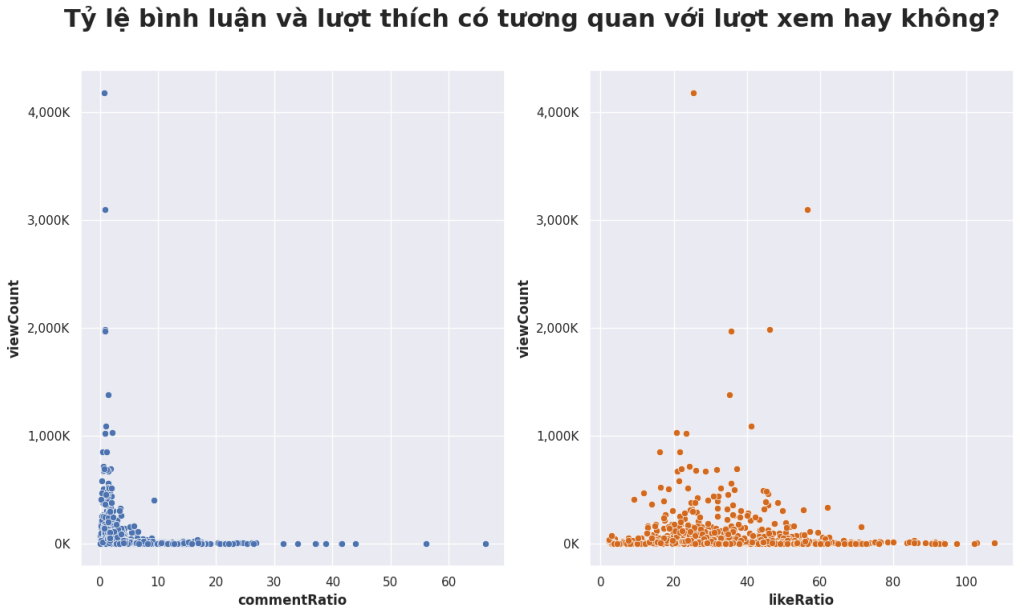


Hình 13. Biểu đồ số lượng views của mỗi kênh youtube

Biểu đồ thể hiện tổng số lượng view của 7 kênh Youtube. Kênh có tổng lượng view cao nhất là Tina Huang với khoảng 20 triệu view. Thấp nhất là kênh The Almost Astrophysicist với khoảng 1 triệu view.

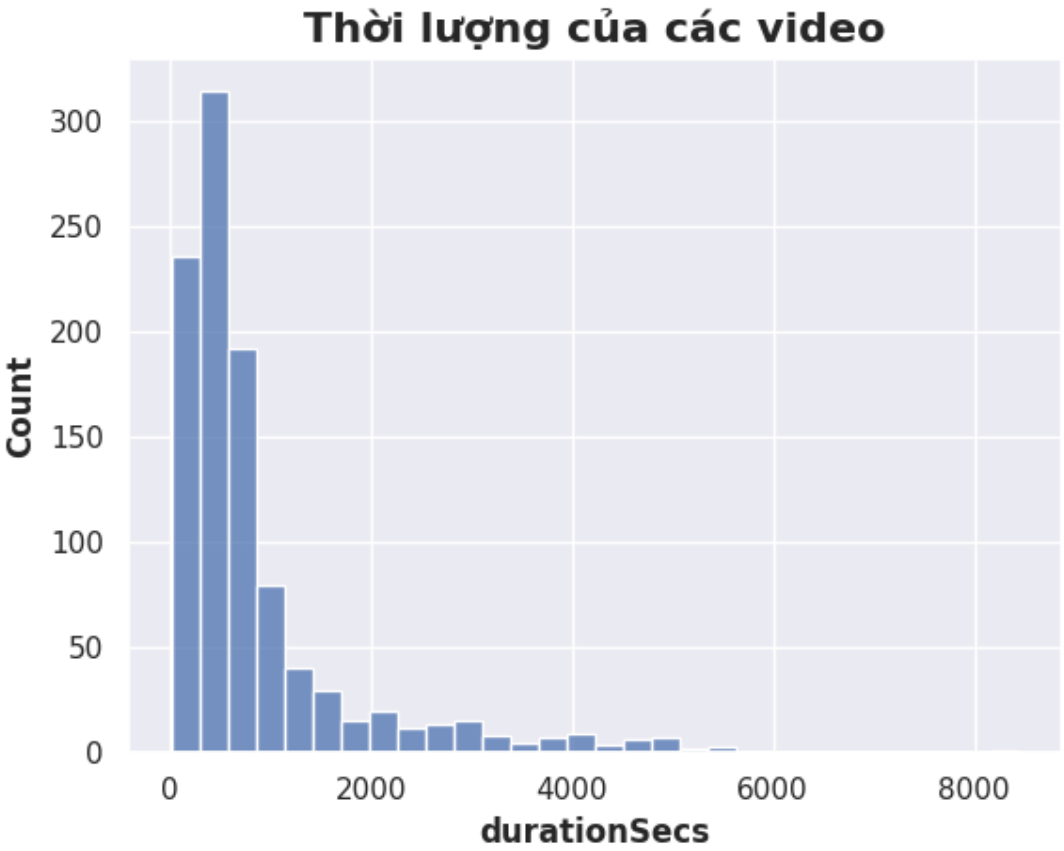


Hình 14. Biểu đồ thể hiện mối quan giữa tổng bình luận, lượt thích với tổng lượt xem



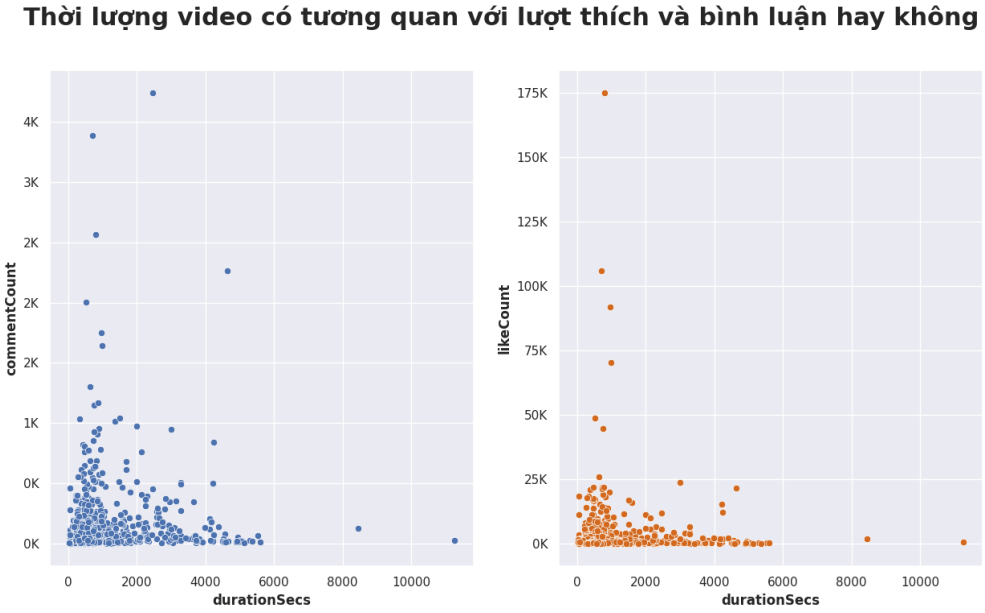
Hình 15. Biểu đồ thể hiện mối quan hệ tỷ lệ giữa bình luận, lượt thích với tổng lượt xem

Với biểu đồ thể hiện tổng lượt bình luận so với lượt xem ta có thể thấy trong những video dưới 1 triệu view thì lượng bình luận dao động trong khoảng 0 → 1000 comment. Ứng với biểu đồ tỷ lệ chiếm khoảng dưới 10%. Ngược lại lượt thích so với lượt xem có vẻ khả quan hơn khi trong khoảng 1 triệu view thì mật độ like dao động từ 0 → 25000 lượt like. Ứng với biểu đồ tỷ lệ nó chiếm khoảng dưới 60%



Hình 16. Biểu đồ thể hiện thời lượng video

Các video thường sẽ có thời lượng trung bình nằm trong khoảng từ 0 → 2000 giây (khoảng 30 phút). Số video trên 4000 giây (hơn 1 giờ) thì chiếm tỷ lệ khá ít.

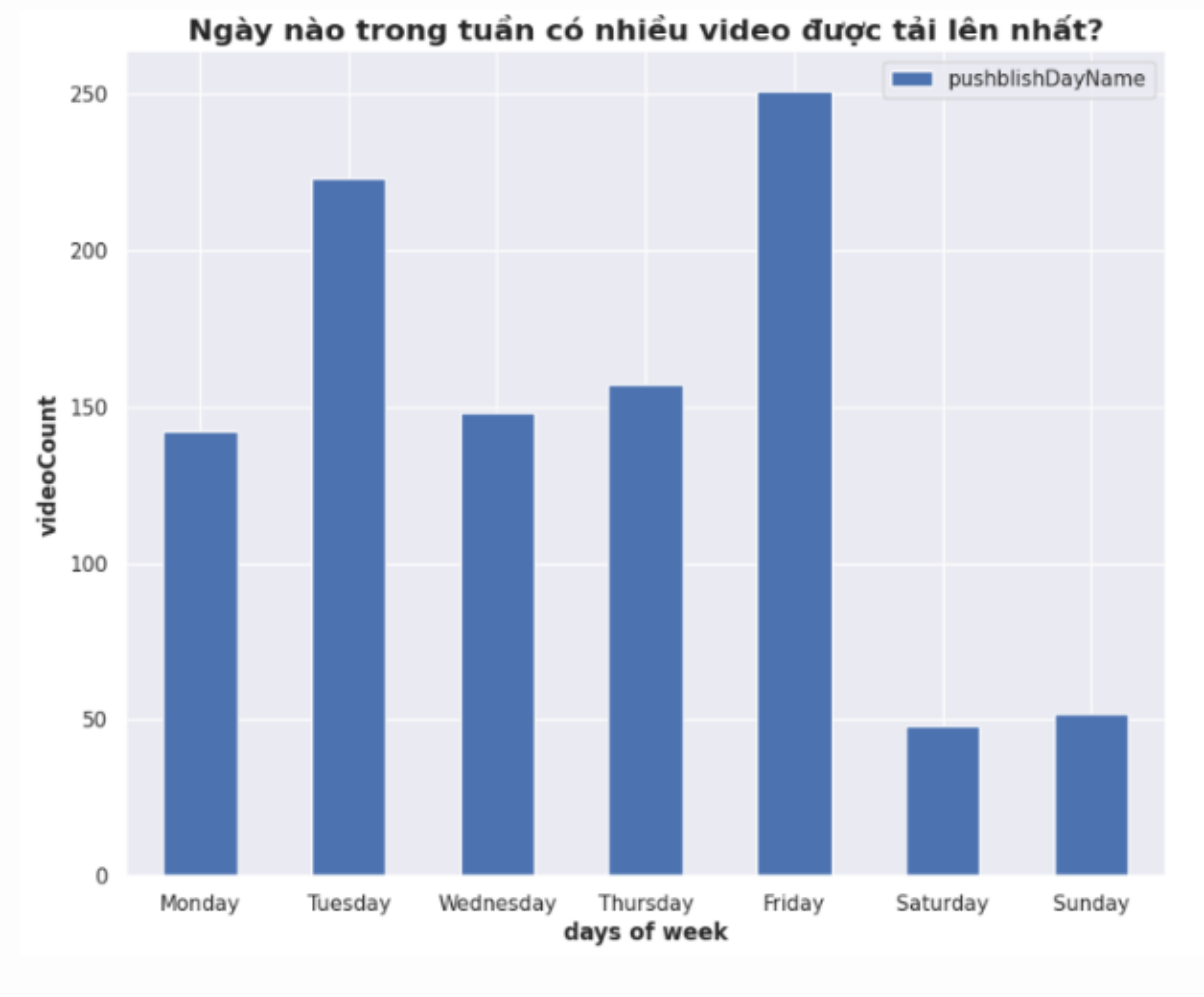


Hình 17. Biểu đồ thể hiện mối quan giữa thời lượng với tổng lượt xem và tổng lượt thích

Nhìn vào biểu đồ này ta thấy từ 0 → 2000 giây, mật độ comment cũng như lượt like tập trung nhiều nhất.

Hình 18. Biểu đồ thể hiện sự tác động của tiêu đề với số lượt xem

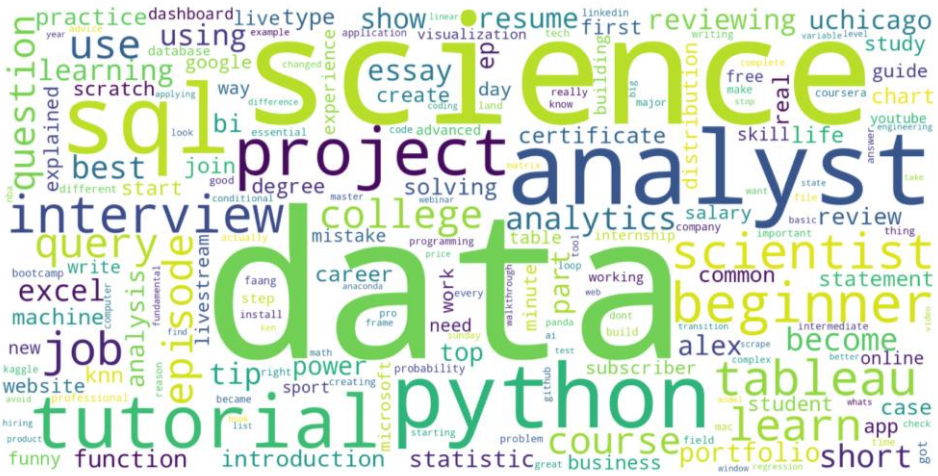
Biểu đồ này thể hiện mối quan hệ tuyến tính không rõ ràng giữa tiêu đề và lượt xem. Dựa vào kết quả nhận được ta chỉ có thể đánh giá được phạm vi mà của tiêu đề dao động trong khoảng 20 → 100 ký tự.



Hình 19. Biểu đồ thể hiện mật độ đăng tải video nhiều nhất

Từ biểu đồ ta nhận thấy các video thường được đăng tải vào các ngày thứ trong tuần (thứ 2 → thứ 6). Video được đăng tải nhiều nhất vào ngày thứ 6 trong tuần, tiếp đến là thứ 3. Các ngày thứ 7 và chủ nhật chiếm tỷ lệ rất thấp.

WordCloud for words in description



Hình 20. WordCloud của biến description

Các từ ngữ thường xuyên xuất hiện trong biến description, phổ biến như: data, science, Python,...Các từ ngữ này đều xoay quanh chủ đề chính của nhóm là topic Data Science.

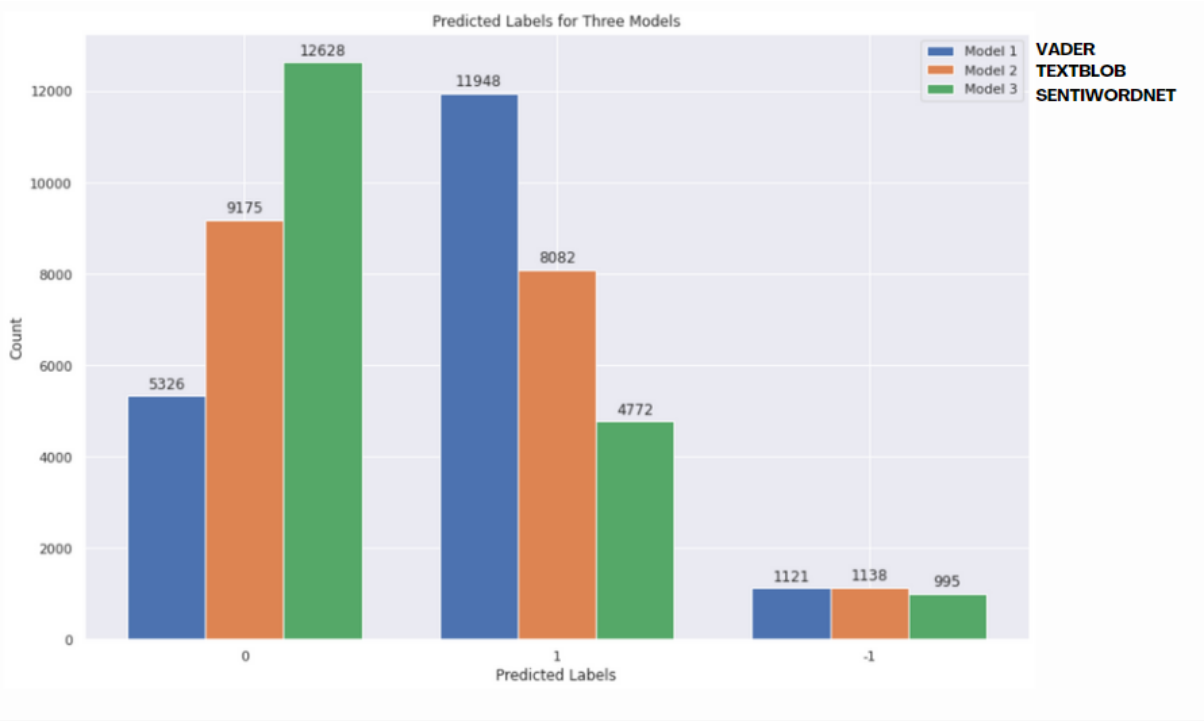
WordCloud for words in comments

Hình 21. WordCloud của biến comments

Đối với biến comment thì xuất hiện những từ ngữ phổ biến như sau: Thank, data, good,...Đa số comment đều ở mức tích cực.

CHƯƠNG 4: KẾT QUẢ và ĐÁNH GIÁ

4.1. Sentiment Analysis



Hình 22. Kết quả đánh giá của các mô hình

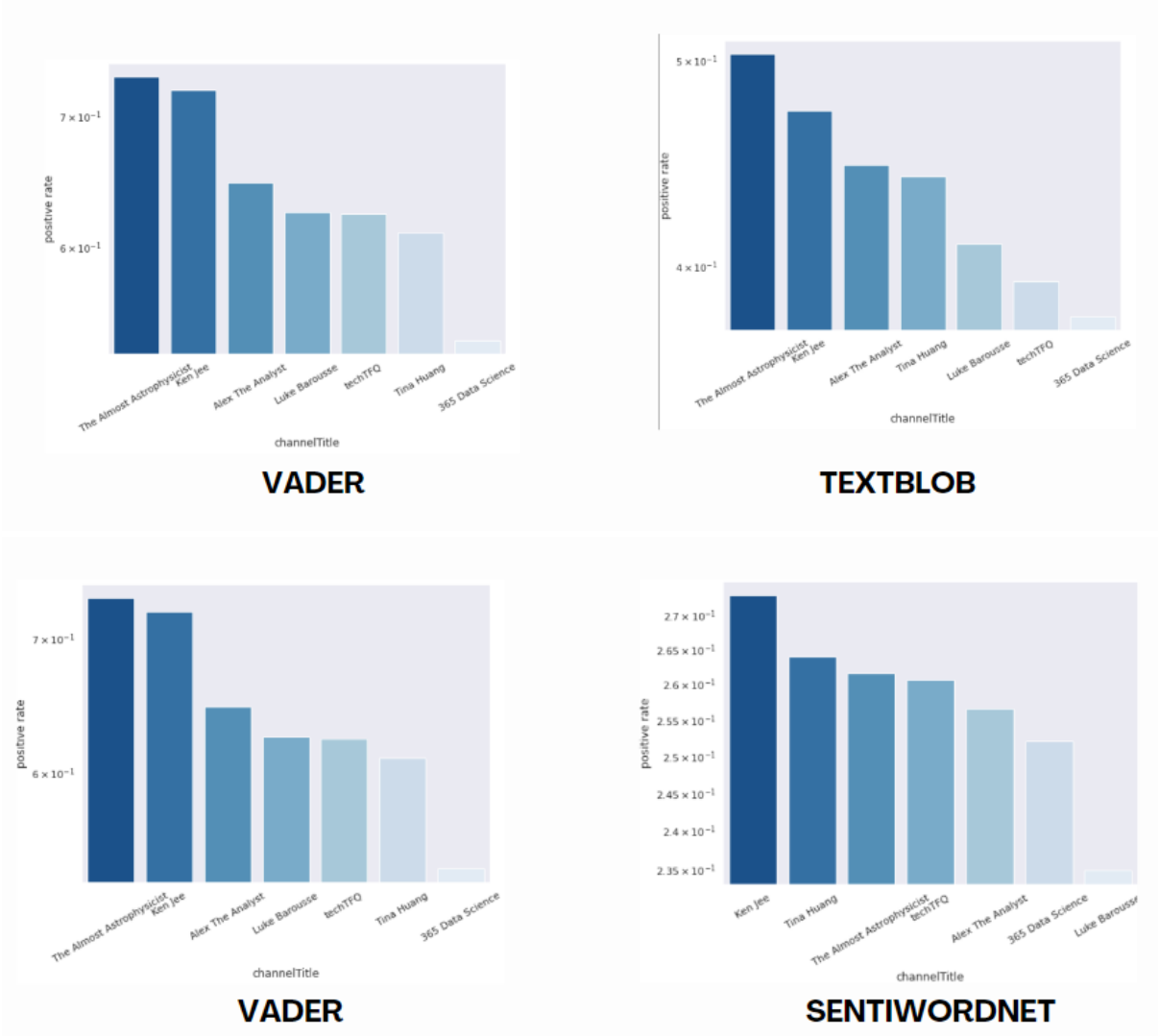
Phân tích:

Nhãn dự đoán *'Negative'* của 3 models có số lượng xấp xỉ nhau với khoảng 995-1138 giá trị.

Nhãn dự đoán *'Positive'* của vader (model 1) là 11948 có số lượng cao hơn khoảng 4000 so với của textblob (model 2) là 8082, và gấp 3 lần con số của SentiWordNet (model 3) là 4772 .

Ngược lại thì, Nhãn dự đoán *'Neutral'* của vader (model 1) thấp nhất trong 3 models, tại 5326, có số lượng thấp hơn mô hình Textblob và SentiWordNet là 9175 và 12628.

4.2. Tìm ra channel có nhiều bình luận tích cực nhất



Hình 23. Channel có nhiều lượt thích nhất qua các mô hình

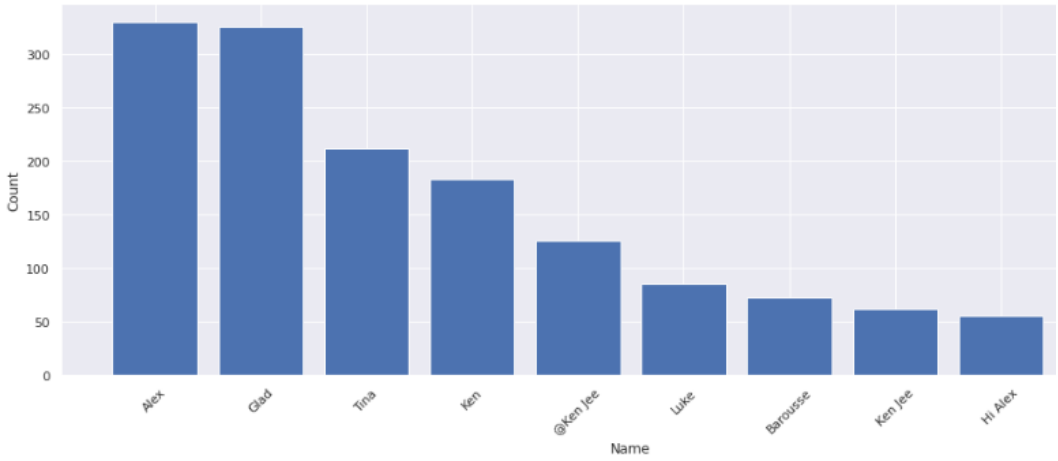
Phân tích:

Ở hai model VADER và TextBlob đều đưa ra kênh *'The Most Astrophysicist'* có tỷ lệ bình luận đánh giá tích cực nhất, và kênh *'365 Data Science'* là thấp nhất. Tuy nhiên, vị trí của kênh *'The Most Astrophysicist'* lại đứng hạng 3 ở model SentiWordNet, thay vào đó là kênh *'Ken Jee'* ở vị trí thứ nhất và *'Luke Barousse'* ở vị trí thấp nhất.

Ngoài ra thì tỷ lệ tích cực của model VADER là cao nhất (0.53-0.73), tiếp theo là textblob(0.37-0.5), cuối cùng là SentiWordNet(0.23-0.27)

4.3. Named Entity Recognition

a. Top các nhân vật xuất hiện



Hình 24. Top các nhân vật thường xuyên xuất hiện

Ở đây ta có thể thấy top các nhân vật có tần suất xuất hiện cao nhất trong 19646 bình luận là: Alex, Glad, Tina, Ken Jee, Luke Barousse.

Vì chủ đề của 7 kênh channel mà nhóm lựa chọn thu thập và phân tích là các kênh về lĩnh vực khoa học dữ liệu, vì vậy top các nhân vật trích xuất bằng phương pháp NER là những người sáng tạo nội dung (content creator) đáng chú ý, làm nội dung về chủ đề phân tích dữ liệu được nhiều người đề cập và quan tâm nhất.

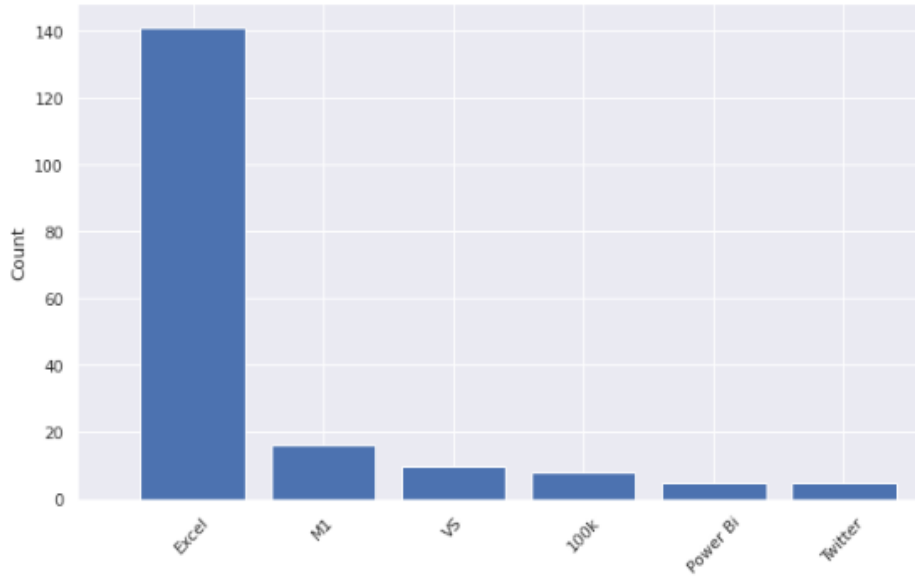
_ Alex: Là người sáng tạo nội dung kênh channel Alex The Analyst, kênh YouTube chuyên đăng tải các video cực kỳ hữu ích về lĩnh vực Data Analytics.

_ Tina: Là người sáng tạo nội dung kênh Tina Huang, kênh YouTube về lĩnh vực Data Science, learning và productivity.

_ Ken Jee: là người sáng tạo nội dung kênh Ken Jee, kênh YouTube làm nội dung về Data Science trong Sport Analytics

_ Luke Barousse: là người sáng tạo nội dung của kênh Luke Barousse, là Data Anlyst làm nội dung trên YouTube về tech và kỹ năng về lĩnh vực phân tích dữ liệu.

b. Top các sản phẩm xuất hiện nhiều nhất



Hình 25. Top các sản phẩm thường xuyên xuất hiện

Top các sản phẩm có tần suất xuất hiện có tần suất xuất hiện cao nhất là Excel, M1, VS, Power BI, Twitter.

Vì 7 kênh YouTube mà nhóm lựa chọn phân tích là các kênh làm về chủ đề khoa học dữ liệu, vì vậy đây là những công cụ được những người quan tâm về lĩnh vực phân tích dữ liệu đề cập nhiều nhất ở trong các bình luận của video:

_ Excel, Power BI là 2 công cụ giúp phân tích và trực quan hóa dữ liệu, xây dựng các biểu đồ, dashboards.

_ M1 ở đây có thể là Macbook Pro M1, một hãng laptop nổi tiếng được khá nhiều nhà phân tích dữ liệu sử dụng vì tính năng mạnh mẽ của nó.

_ VS ở đây là trong VS Code (Visual Studio Code), trình soạn thảo, biên tập lập trình mã nguồn miễn phí được các nhà khoa học dữ liệu, nhà phân tích dữ liệu sử dụng cực kỳ phổ biến vì sự tiện dụng của nó.

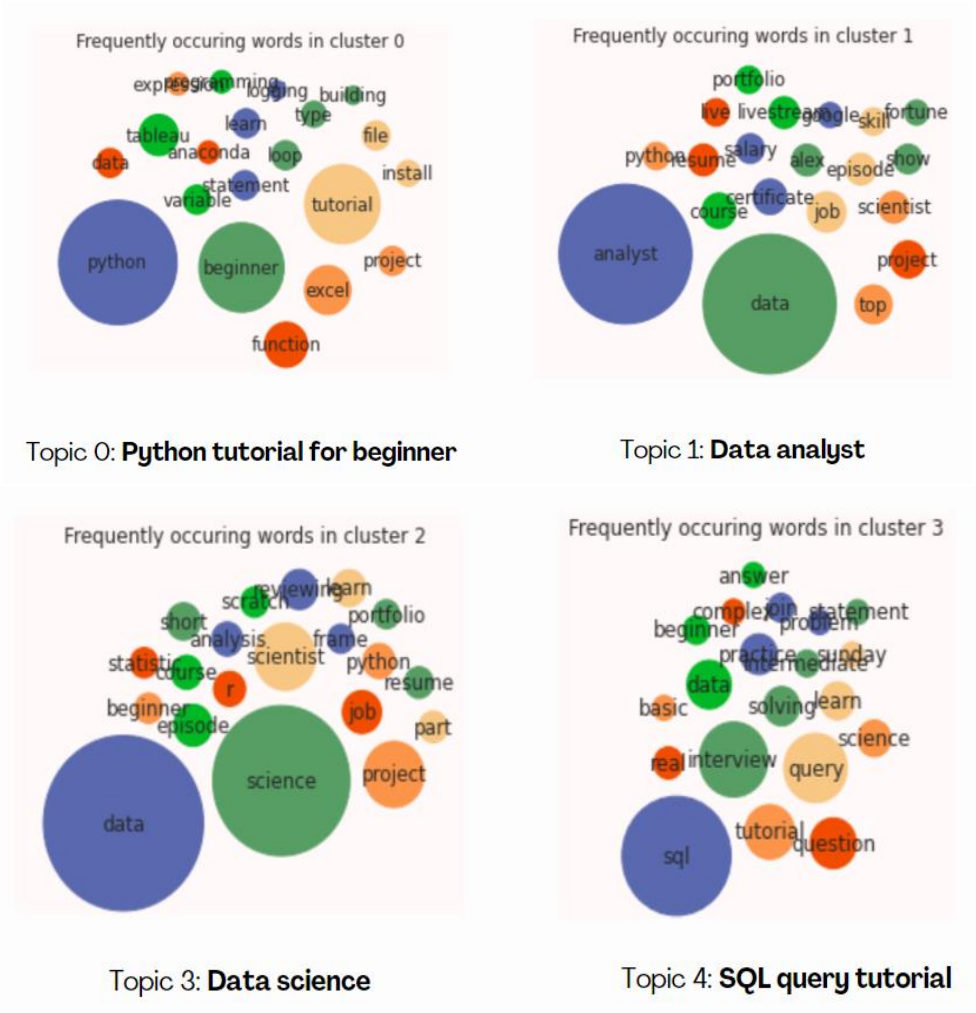
_ Twitter là một trang mạng xã hội cực kỳ phổ biến và hỗ trợ các nhà phát triển (developer) truy cập vào dữ liệu Twitter thông qua API Key. Những nhà khoa học dữ liệu thường sử dụng dữ liệu Twitter trong việc thu thập dữ liệu mạng xã hội để phân tích, áp dụng trong các bài toán xử lý ngôn ngữ tự nhiên. Vì vậy đó là lý do vì sao Twitter lọt top những công cụ được đề cập nhiều nhất trong các bình luận của các kênh YouTube này.

4.4. Topic Extraction (K-Means)

Thực hiện phân cụm bằng K-Means để trích xuất các chủ đề nổi bật trong 1079 videos của 7 kênh channel về lĩnh vực khoa học dữ liệu

Ở đây ta trích xuất ra các chủ đề sau:

+ Topic 0: Có thể nói rằng những video chủ yếu nhắm tới những người đang bắt đầu học và tìm hiểu lĩnh vực khoa học dữ liệu, và ngôn ngữ lập trình được những nhà phân tích dữ liệu, nhà khoa học dữ liệu quan tâm và sử dụng nhiều nhất là Python, với những keyword: 'python', 'beginner', 'tutorial'.
+ Topic 1: Chủ đề nói đến những đặc điểm về công việc Data Analyst, trong đó 'data' và 'analyst' được xuất hiện nhiều nhất, sau đó là 'salary', 'project', 'resume'



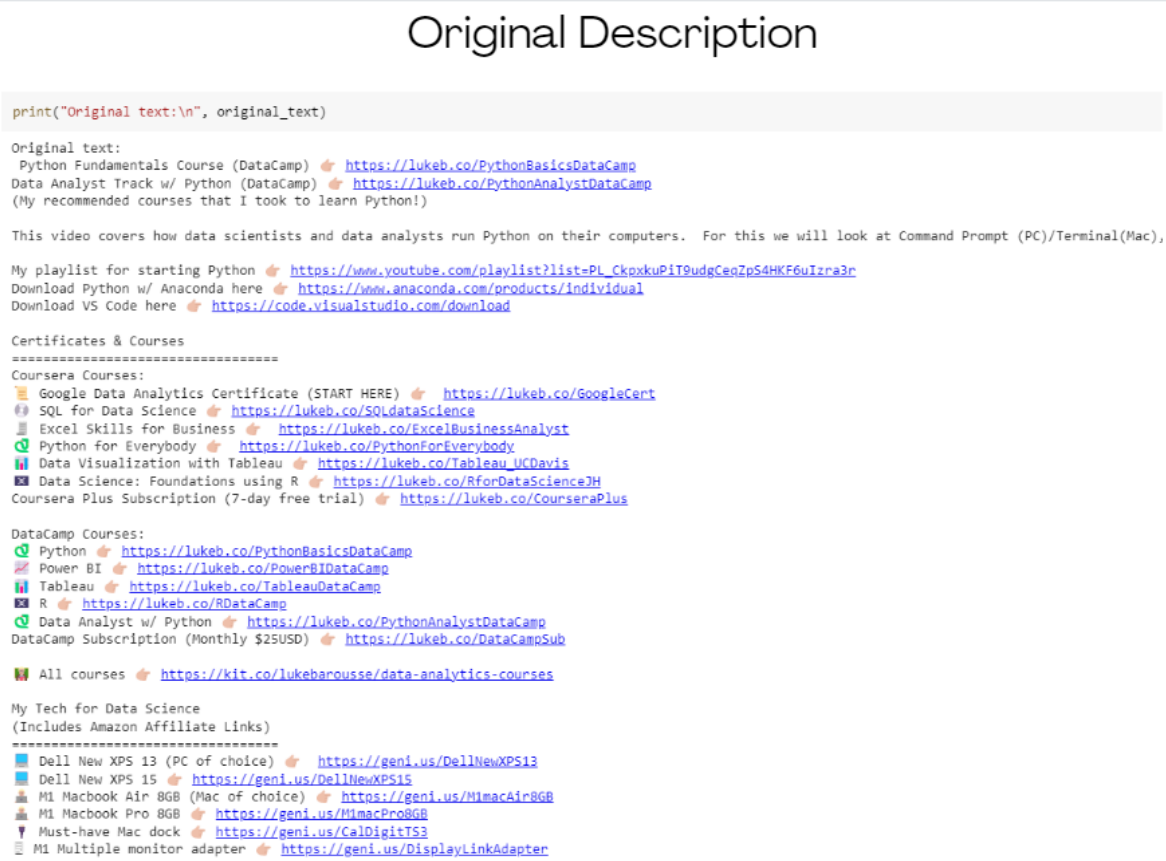
Hình 26. Kết quả về các topic thu được

+ Topic 3: Trong lĩnh vực phân tích dữ liệu thì Data Science là một ngành được rất nhiều sự quan tâm chú ý. Vì vậy các nhà sáng tạo nội dung đã thường xuyên đăng tải những video về chủ đề này. Liên quan đến các 'project' trong lĩnh vực khoa học dữ liệu, ngôn ngữ lập trình phổ biến trong data science như 'Python', 'R'. Video về cách ứng tuyển cho vị trí về data science 'job', 'resume'

+ Topic 4: Đến với chủ đề này thì từ ngữ ‘SQL’, ‘query’ được xuất hiện nhiều nhất, thì có thể rằng đây cũng là ngôn ngữ truy cấn cơ sở dữ liệu được sử dụng phổ biến trong lĩnh vực khoa học dữ liệu. Ngoài ra thì từ khóa ‘interview’ xuất hiện khá nhiều. cho thấy các video về chủ đề về quá trình phỏng vấn với các câu hỏi liên quan đến kỹ năng sử dụng SQL cũng được các nhà sáng tạo nội dung tập trung sản xuất video.

4.5. BERT Summarization

Sử dụng mô hình Summarizer (BERT-based summarization model) để thực hiện tóm tắt văn bản đối với biến ‘Description’.
Có thể thấy phần mô tả của video trước khi được làm sạch và tóm tắt là rất dài, bao gồm các thông tin ít thể hiện nội dung chính của video, đa số là các đường dẫn và thông tin tới các khóa học mà chủ kênh nhận quảng bá (Coursera Courses, DataCamp Courses). Các link này hầu như xuất hiện ở hầu hết các video mà chủ kênh đăng tải, và không thể hiện nội dung chính mà video muốn truyền tải.

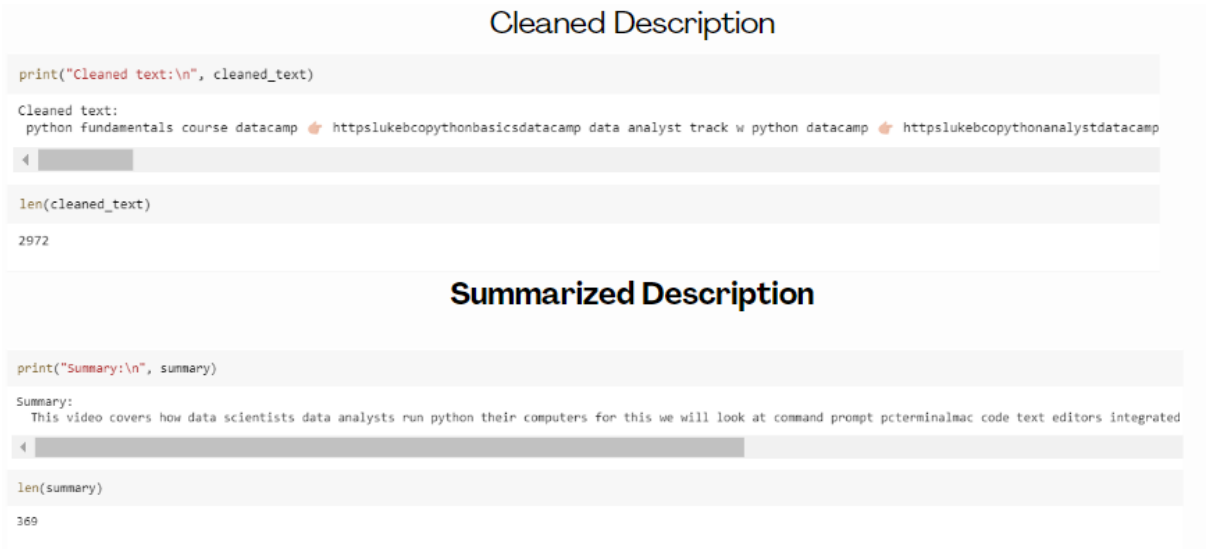


Hình 27. Biến ‘Description’ gốc trước khi làm sạch và tóm tắt

Có thể thấy rõ sau khi tiến hành làm sạch và tóm tắt văn bản bằng phương pháp BERT Extractive Summarizer của bộ thư viện PyPI, phần mô tả video đã thể hiện được nội dung chính của video, giúp người xem dễ dàng xác định được nội dung

chính mà video truyền tải là về những vấn đề nào, giúp người xem tiết kiệm thời gian hơn so với việc lướt qua một Description chưa qua làm sạch và tóm tắt bao gồm nhiều thông tin quảng cáo khóa học.

Sau khi tóm tắt thì độ dài của Description đã giảm từ 2972 ký tự xuống còn 269 ký tự, thể hiện rõ ràng được nội dung chính của video, mạch văn cũng rất mượt và dễ hiểu, không bị quá vắn tắt khiến cho người đọc khó đọc hiểu được nội dung của văn bản.



Hình 28. Biến 'Description' sau khi summarized

```
Summary:
 This video covers how data scientists data analysts run python their computers for this we will look at command prompt pcterminalmac code text editors integrated development environments ides jupyter notebooks for all these examples we look at how use them their use case recommended options my playlist for starting python is for python fundamentals course datacamp .
```

Như ví dụ trên, ta có thể nhanh chóng xác định nội dung chính của video là trình bày cách các nhà khoa học dữ liệu và nhà phân tích dữ liệu sử dụng công cụ Python. Bản mô tả video sau khi tóm tắt đã lược bớt các thông tin quảng cáo khóa học dày đặc trong văn bản. Giúp người đọc nhanh chóng và dễ dàng xác định được nội dung video mà họ quan tâm.

CHƯƠNG 5: KẾT LUẬN

1. Nhận xét

Trong dự án này, nhóm đã khám phá dữ liệu video của 7 kênh Khoa học dữ liệu/Phân tích dữ liệu tương đối phổ biến và tiết lộ nhiều phát hiện thú vị cho bất kỳ ai bắt đầu với kênh Youtube về khoa học dữ liệu hoặc chủ đề khác:
Video càng có nhiều lượt thích và bình luận thì video đó càng nhận được nhiều lượt xem (nó thể hiện một mối tương quan và có thể tác động theo cả hai chiều). Lượt thích dường như là một chỉ số tương tác tốt hơn bình luận và số lượt thích dường như tuân theo "bằng chứng xã hội", có nghĩa là video càng có nhiều lượt xem thì càng có nhiều người thích video đó.
Các video được xem nhiều nhất thường có độ dài tiêu đề trung bình từ 30-80 ký tự. Tiêu đề quá ngắn hoặc quá dài dường như gây hại cho lượng người xem.
Video thường được tải lên vào thứ Hai và thứ Sáu. Đặc biệt, cuối tuần và Chủ nhật không phải là thời điểm phổ biến để đăng video mới.
Các nhận xét về video nhìn chung là tích cực, nhóm nhận thấy rất nhiều từ "Thank" và “good”, gợi ý những khoảng trống thị trường tiềm ẩn trong nội dung có thể được lấp đầy. Điều này có thể nhìn rõ ở phần phân tích cảm xúc bằng việc dùng VADER, một công cụ phân tích đáng tin cậy trong các văn bản mạng xã hội.

2. Hướng phát triển

Để mở rộng và xây dựng dựa trên dự án nghiên cứu này, người ta có thể:
Mở rộng tập dữ liệu sang các kênh nhỏ hơn trong lĩnh vực khoa học dữ liệu
Thực hiện phân tích cảm tính đối với các nhận xét và tìm ra video nào nhận được nhiều nhận xét tích cực hơn và video nào nhận được ít nhận xét tích cực hơn.
Thực hiện nghiên cứu thị trường bằng cách phân tích các câu hỏi trong chuỗi nhận xét và xác định các câu hỏi phổ biến/khoảng trống thị trường có khả năng lấp đầy.
Tiến hành nghiên cứu này cho các thị trường ngách khác (ví dụ: vlog hoặc kênh làm đẹp), để so sánh các thị trường ngách khác nhau với nhau nhằm xem các mẫu khác nhau về lượng người xem và đặc điểm video.

TÀI LIỆU THAM KHẢO

[1], 2020. *5 YouTube Algorithm Myths YouTubers NEED to Know With Mark Robertson.* [Online]
Available at: https://vidiq.com/vi/blog/post/5-youtube-algorithm-myths-youtubers-need-to-know-about/
[2], 2019. *Big Data là gì? 5 Vs của Big Data là gì? Tại sao nó quan trọng?.* [Online]
Available at: https://viblo.asia/p/big-data-la-gi-5-vs-cua-big-data-la-gi-tai-sao-no-quan-trong-gAm5yWALZdb
[3], 2020. *5V - Bigdata.* [Online]
Available at: https://aramex.vn/khai-thac-big-data-duoc-dien-ra-cu-the-nhu-the-nao.html/big-data-5/
[4], 2018. *Các giai đoạn tiền xử lý.* [Online]
Available at: https://blog.vietnamlab.vn/ban-ve-cong-doan-tien-xu-ly-trong-xu-ly-ngon-ngu-tu-nhien/
[5], 2020. *TF-IDF là gì? Code demo thuật toán TF-IDF với dữ liệu tiếng Việt.* [Online]
Available at: https://sentayho.com.vn/tf-idf-la-gi.html
[6], 2021. *Sentiment analysis.* [Online]
Available at: https://en.wikipedia.org/wiki/Sentiment_analysis
[7], 2022. *VADER-Sentiment-Analysis.* [Online]
Available at: https://github.com/cjhutto/vaderSentiment#vader-sentiment-analysis
[8], 2020. *View of VADER.* [Online]
Available at: https://ojs.aaai.org/index.php/ICWSM/article/view/14550/14399
[9], 2010. *SENTIWORDNET 3.0: AnEnhancedLexicalResource forSentimentAnalysisandOpinionMining.* [Online]
Available at: http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf
[10], 2021. *Sentiment Analysis with Textblob and Vader in Python.* [Online]
Available at: https://www.analyticsvidhya.com/blog/2021/10/sentiment-analysis-with-textblob-and-vader/#:~:text=As%20TextBlob%20is%20a%20Lexicon,Rule%2Dbased%20sentiment%20analyzers%E2%80%9D
[11], 2020. *Named-entity recognition - Wikipedia.* [Online]
Available at: https://en.wikipedia.org/wiki/Named-entity_recognition
[12], 2021. *Phân cụm k-Means.* [Online]
Available at: https://aiwithmisa.com/2021/06/14/aml-bai17/
[13], 2019. *k-Means clustering.* [Online]
Available at: http://edu.sablab.net/rp2019/scripts4.html
[14], 2018. *Khái niệm BERT.* [Online]
Available at: https://viblo.asia/p/bert-buoc-dot-pha-moi-trong-cong-nghe-xu-ly-ngon-ngu-tu-nhien-cua-google-RnB5pGV7lPG
[15], 2022. *BERT for text summarization.* [Online]
Available at: https://ig.opengenus.org/bert-for-text-summarization/