Kỷ yếu Hội nghị Khoa học công nghệ Quốc gia lần thứ XVI về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Đà Nẵng, ngày 28-29/9/2023
DOI: 10.15625/vap.2023.0049

MỘT TIẾP CẬN TÓM TẮT NỘI DUNG TRONG VIDEO TIN TỨC

Nguyễn Thanh Hải[1], Lê Việt Khoa[1,2], Đỗ Khánh Toàn[1], Nguyễn Thái Nghe[1]

[1]Trường Công nghệ thông tin và Truyền thông, Trường Đại học Cần Thơ
[2]Đài Phát thanh và Truyền hình thành phố Cần Thơ

nthai.cit@ctu.edu.vn, vietkhoa@thtpct.vn, ktoan515@gmail.com, ntnghe@cit.ctu.edu.vn

TÓM TẮT: *Với sự bùng nổ của Internet, nguồn thông tin ngày càng nhiều và đa dạng dẫn đến nhu cầu phân loại và tóm tắt thông tin ngày càng trở nên vô cùng cần thiết. Những bản tin từ các video là nguồn thông tin rất có giá trị giúp chúng ta cập nhật những bản tin có được thông tin thời sự nhất. Tuy nhiên, với những video khá dài, trong khi cuộc sống ở các đô thị diễn ra bận rộn gây khó khăn cho người dùng để có thể theo dõi hết toàn bộ video. Đồng thời nhu cầu tìm kiếm nội dung từ video cũng rất cần thiết. Tóm tắt các nội dung video có thể giúp người dùng nắm bắt nhanh các chủ đề và nội dung video, đồng thời hỗ trợ cho các công cụ tìm kiếm để có được những thông tin mà người dùng quan tâm thay vì phải xem hết toàn bộ nội dung video.*

Trong nghiên cứu này, chúng tôi đánh giá độ hiệu quả của 6 phương pháp tóm tắt văn bản trên các video tập hợp từ 9 chủ đề tin tức phổ biến. Trước khi tóm tắt, chúng tôi đã tiến hành lọc bớt nhiễu để từ đó có nội dung văn bản được trích xuất tốt hơn. Kết quả thực nghiệm cho thấy Textrank cho kết quả tốt nhất và các Video đưa tin tức về thế giới, và bản tin Khoa học cho độ chính xác tóm tắt cao nhất. Kết quả cũng cho thấy việc lọc nhiễu là rất cần thiết khi trích xuất nội dung văn bản trong các video tin tức. Phương pháp đề xuất mong đợi có thể tiếp tục cải tiến để áp dụng trong các lĩnh vực xử lý tóm tắt nội dung áp dụng cho các đài truyền hình, đưa tin tức phóng sự hàng ngày.

Từ khóa: *Lọc nhiễu, Tóm tắt văn bản, Video tin tức.*

I. GIỚI THIỆU

Với sự phát triển không ngừng của internet, số lượng video trên mạng ngày càng tăng, đòi hỏi phải tổ chức và phân loại chúng hiệu quả dựa trên nội dung. Trong việc phân loại các chủ đề của video, việc tự động nhận dạng các chủ đề chính trong video là một khía cạnh quan trọng. Các tác vụ này có nhiều ứng dụng tiềm năng để tìm và khám phá nội dung video cũng như hỗ trợ phân tích lượng lớn dữ liệu video cho mục đích nghiên cứu hoặc quảng cáo. Việc phân loại theo chủ đề cũng thuận tiện cho việc phân tích lượng lớn dữ liệu video nhằm mục đích nghiên cứu xu hướng hoặc mẫu hành vi của người dùng cũng như có thể sử dụng để phân loại video trong truyền hình, phục vụ cho công tác sản xuất, lưu trữ và phân phối. Bên cạnh việc phân loại theo chủ đề, nhu cầu tóm tắt xử lý văn bản dưới dạng video cũng ngày càng nhiều. Điều này bao gồm các tác vụ như trích xuất văn bản có liên quan từ video, tạo chú thích và phụ đề cũng như thực hiện phân tích cảm xúc trong văn bản. Các tác vụ này có nhiều ứng dụng tiềm năng, bao gồm cải thiện khả năng tìm kiếm và khả năng khám phá nội dung video, cung cấp khả năng tiếp cận tốt hơn cho những người khuyết tật về thị giác hoặc thính giác và hỗ trợ phân tích lượng lớn dữ liệu video cho mục đích nghiên cứu, khai thác, lưu trữ và phân phối hiệu quả các video trên môi trường mạng hoặc các nền tảng video khác.

Trong nghiên cứu này, chúng tôi đánh giá một số phương pháp để tóm tắt văn bản được trích xuất từ các video tin tức. Chúng tôi thảo luận về một số hướng nghiên cứu về tóm tắt văn bản và đánh giá những tiềm năng trong lĩnh vực này trên các nội dung văn bản được trích xuất từ các Video. Chúng tôi tiến hành xử lý video trước khi trích xuất nội dung văn bản bằng các thuật toán lọc nhiễu để có thể trích xuất được nhiều nội dung văn bản thông qua giọng nói phát thanh viên. Sau đó thực hiện các phương pháp để tóm tắt nội dung văn bản từ video và đánh giá so sánh giữa giải thuật được khảo sát, phân tích các kết quả tóm tắt giữa các chủ đề. Như đã thể hiện trong kết quả thực nghiệm, đóng góp chủ yếu của nghiên cứu gồm đánh giá phân tích tiềm năng của các kỹ thuật xử lý nhiễu âm thanh, tóm tắt văn bản để thể hiện tính ứng dụng thực tiễn của các thuật toán này trong việc tóm tắt nội dung trong các video tin tức.

II. CÁC NGHIÊN CỨU CÓ LIÊN QUAN

Các nghiên cứu xử lý văn bản trong video là một chủ đề được quan tâm, với nhiều nghiên cứu được đề xuất trong những năm gần đây. Một khía cạnh quan trọng trong trích xuất nội dung từ các giọng nói (từ phát thanh viên) là giảm tiếng ồn nhằm mục đích loại bỏ bất kỳ yếu tố ngoại lai có thể ảnh hưởng trích xuất nội dung từ giọng nói trong video, có một số đề xuất giảm tiếng ồn khi nhận dạng giọng nói trong [1, 2]. Nội dung video cũng có thể trích xuất từ hình ảnh từ các khung hình trong video nhưng khá thử thách do phông chữ, màu sắc và nền của hình ảnh, cũng như sự hiện diện của các yếu tố hình ảnh khác có thể cản trở quá trình nhận dạng văn bản.

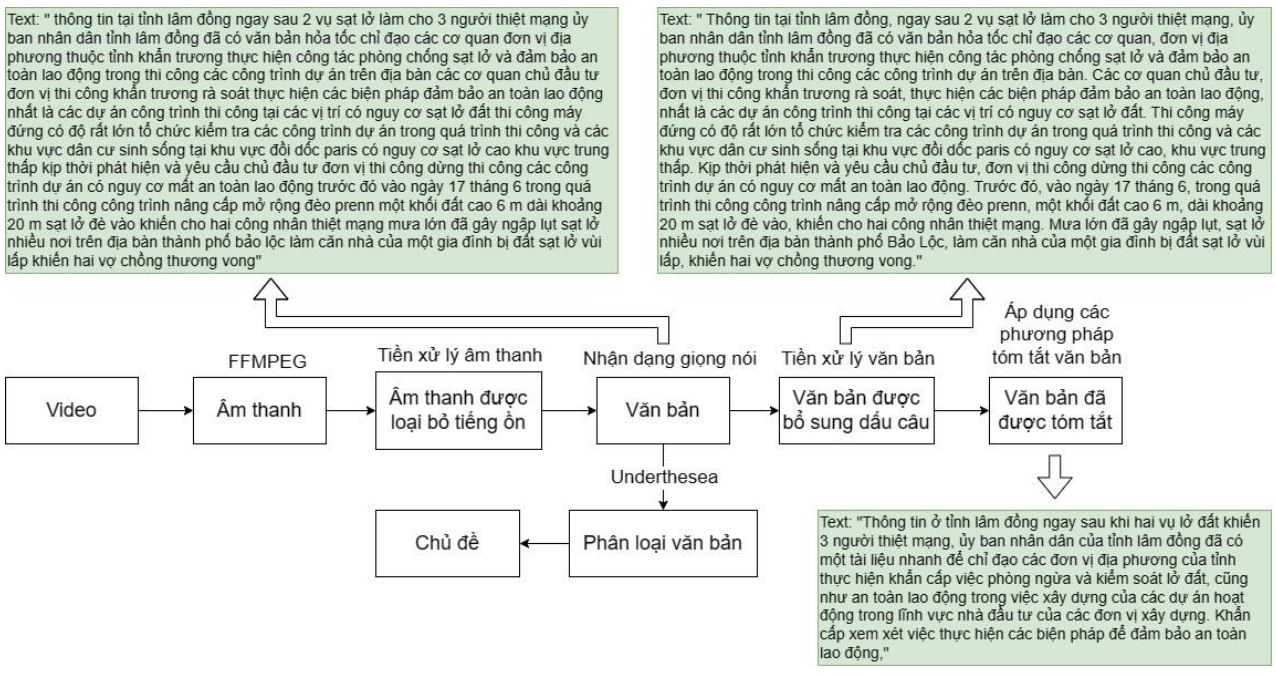
Thách thức quan trọng trong xử lý văn bản cho video là tóm tắt văn bản gồm các tác vụ xác định thông tin quan trọng cô đọng nhất trong video. Nhiều kỹ thuật khác nhau đã được đề xuất cho mục đích này, bao gồm xử lý ngôn ngữ tự nhiên, học máy và học sâu. Các phương pháp này thường liên quan đến việc phân tích nội dung âm thanh, hình ảnh và văn bản của video để xác định các chủ đề và khái niệm chính, sau đó tạo một bản tóm tắt chứa thông tin cần thiết. Trong các nghiên cứu trước đây về tóm tắt văn bản tự động, các nhà nghiên cứu đã thử nghiệm một loạt các mô hình máy học để tạo ra các bản tóm tắt chính xác. Chẳng hạn, một nghiên cứu ở [3] đã sử dụng bảy mô hình cơ sở học máy khác nhau, bao gồm Hồi quy logistic, Cây quyết định, Rừng ngẫu nhiên, Naïve Bayes, XGBoost, Mạng nơron thần

kinh và SVM để tạo bản tóm tắt. Sau đó, các tác giả đã sử dụng phương pháp học đồng bộ để kết hợp các dự đoán từ cả bảy mô hình, dẫn đến dự đoán chính xác hơn bất kỳ mô hình riêng lẻ nào có thể đạt được. Các thông số sử dụng trong kỹ thuật này được lựa chọn cẩn thận để đạt được độ chính xác tối đa sau nhiều lần lặp lại. Để cải thiện hơn nữa kết quả, các tác giả đã sử dụng các kỹ thuật học cụm, cụ thể là sử dụng phương pháp phân cụm k-mean để nhóm các câu tương tự lại với nhau. Cách tiếp cận này đã cải thiện tính mạch lạc và dễ đọc của các bản tóm tắt. Các tác giả đã đạt được kết quả tốt nhất khi sử dụng các mô hình Rừng ngẫu nhiên, XGBoost và Mạng thần kinh, với mô hình của họ đạt được điểm ROUGE-1 là 0,78, điểm ROUGE-2 trung bình là 0,66 và điểm ROUGE-L là 0,76. Thêm nữa, hầu hết các phương pháp chấm điểm câu được sử dụng là phổ biến. Các phương pháp chấm điểm được phân loại như chấm điểm từ, chấm điểm câu và chấm điểm đồ thị điển hình như [4]. Bộ dữ liệu trong các nghiên cứu đó bao gồm các bài báo y sinh và các câu ứng cử viên được trích xuất để tóm tắt bằng cách sử dụng phương pháp dựa trên quy tắc. Sau đó, chúng biểu diễn các câu này và các mối quan hệ của chúng dưới dạng biểu đồ, với các nút biểu thị các câu và các cạnh biểu thị sự giống nhau về ngữ nghĩa giữa các câu. Họ đã sử dụng khai thác tập mục để xác định các cụm từ quan trọng trong câu và sử dụng các cụm từ này để nhóm các câu thành các nhóm mạch lạc. Kết quả của nghiên cứu cho thấy rằng phương pháp đề xuất của họ vượt trội hơn một số phương pháp cơ bản về cả điểm số ROUGE và đánh giá của con người. Mặc dù những nghiên cứu này đã cho thấy kết quả đầy hứa hẹn trong lĩnh vực y sinh, nhưng vẫn chưa rõ liệu phương pháp đề xuất có thể được mở rộng sang các lĩnh vực khác như phương tiện truyền thông xã hội và tin tức hay không. Trong nghiên cứu từ [5], các tác giả đã tóm tắt văn bản bằng cách xếp hạng các câu và các câu được chọn vào phần tóm tắt. Một phương pháp khác đề xuất cách tiếp cận theo ngữ nghĩa [6]. Để đạt được một bản tóm tắt chính xác nhất, các tác giả trong [6] đã loại bỏ những câu văn ngắn và lọc bỏ những thông tin không liên quan. Mô hình được đề xuất tạo ra các câu theo trình tự thời gian dựa vào thứ tự chúng xuất hiện trong tài liệu gốc và sử dụng phương pháp tiếp cận heuristic để chọn nhóm câu có ý nghĩa tốt nhất cho bản tóm tắt. Các tác giả cho rằng tiếp cận của họ khắc phục được những thiếu sót của các phương pháp hiện có và cung cấp một bản tóm tắt cân bằng hơn, và họ cũng sử dụng các mô hình tóm tắt văn bản như các kỹ thuật dựa trên phương pháp Luhn [7], Edmonson [8], Latent Semantic Analysis (LSA) [9], LexRank [10] và TextRank [11] để so sánh trong các thực nghiệm trên các tài liệu văn bản thu thập từ Wikipedia. Kết quả thể hiện LexRank cao hơn TextRank, và các phương pháp dựa trên LSA, Luhn, Edmonson. Dù vậy, cách biệt chính xác giữa LexRank và TextRank là không lớn. Trong nghiên cứu này, chúng tôi sẽ tiến hành thực nghiệm để đánhg giá các giải thuật này trên các nội dung văn bản được trích xuất từ giọng nói trong các video tin tức với 9 chủ đề khác nhau. Các giải thuật này vẫn được sử dụng trong các nghiên cứu gần đây để xử lý và tóm tắt văn bản [12, 13].

III. PHƯƠNG PHÁP THỰC HIỆN

Trong phần này, chúng tôi sẽ trình bày phương pháp được đề xuất để tóm tắt nội dung văn bản từ các giọng nói trong video bao gồm các bước chính như minh họa trong Hình 1: lọc nhiễu, trích xuất nội dung từ các giọng nói, phân loại và tóm tắt nội dung âm thanh trong video.

Để chuyển đổi video thành âm thanh, chúng tôi sử dụng FFmpeg [14] để trích xuất âm thanh từ tệp video gốc. Tiếp đó, chúng tôi áp dụng hai phương pháp để loại bỏ tiếng ồn khỏi âm thanh đã được chuyển đổi. Phương pháp đầu tiên là Noisereduce[15, 16], giúp giảm tiếng ồn trong âm thanh và phương pháp thứ hai là DeepFillterNet [17], một mô hình học sâu được sử dụng để loại bỏ tiếng ồn khỏi âm thanh.



Hình 1. Minh họa luồng công việc đề xuất để phân tích, phân loại và tóm tắt văn bản từ video

Sau khi loại bỏ tiếng ồn, chúng tôi chuyển đổi âm thanh đã xử lý thành văn bản với SpeechRecognition [18]. Sau khi giảm nhiễu từ các tiếng ồn, văn bản được trích xuất có vẻ gần hơn với văn bản từ giọng nói trong video gốc. Tiếp theo, chúng tôi thực hiện phân loại theo chủ đề của văn bản được tạo ra để xác định chủ đề hoặc lĩnh vực của nội dung văn bản. Sau khi phân loại các chủ đề, chúng tôi sử dụng Punctuation 2 [19] để xử lý văn bản bằng cách thêm dấu câu. Điều này cải thiện khả năng đọc và hiểu văn bản, tạo ra một phiên bản đáng tin cậy hơn. Sau đó chúng tôi kiểm tra chính tả và phân đoạn câu của văn bản. Thông qua quá trình này, chúng tôi sửa lỗi chính tả và chia văn bản thành các câu riêng biệt, cải thiện độ chính xác và khả năng đọc của văn bản. Cuối cùng, chúng tôi sử dụng các thuật toán tóm tắt để tạo ra một bản tóm tắt hoàn chỉnh. Thông qua tóm tắt, chúng tôi trích xuất thông tin chính và ý chính từ văn bản, tạo ra một phiên bản ngắn gọn nhưng chính xác của nội dung gốc.

A. Kỹ thuật loại bỏ tiếng ồn

Noisereduce [15, 16] là một thuật toán giảm nhiễu dựa trên Python giúp giảm nhiễu trong các tín hiệu miền thời gian, chẳng hạn như lời nói, âm sinh học và tín hiệu sinh lý. Nó sử dụng một kỹ thuật gọi là "ngưỡng quang phổ", là một dạng của cổng nhiễu. Thuật toán tính toán biểu đồ phổ của tín hiệu (cụ thể là nhiễu), ước tính ngưỡng nhiễu cho dải tần số của từng tín hiệu và sử dụng nó để tính toán mặt nạ. Bộ lọc tiếng ồn này rất quan trọng trong việc ngăn tiếng ồn giảm xuống dưới ngưỡng thay đổi tần số. Noisereduce bao gồm 2 cách tiếp cận:

Giảm tiếng ồn cố định (Stationary Noise Reduction). Giữ ngưỡng tiếng ồn ước tính ở cùng mức trên toàn bộ tín hiệu. Thuật toán này nhận 2 tham số đầu vào là âm thanh nhiễu có chứa nguyên mẫu nhiễu (tiếng ồn) và âm thanh tín hiệu có chứa tín hiệu và tiếng ồn dự định được loại bỏ. Các bước của thuật toán giảm tiếng ồn cố định gồm:

1. Tính toán quang phổ trên clip âm thanh tiếng ồn.
2. Số liệu thống kê được tính toán trên phổ của tiếng ồn (theo tần số).
3. Ngưỡng được tính toán dựa trên số liệu thống kê về nhiễu.
4. Một quang phổ được tính toán trên tín hiệu.
5. Bộ lọc được xác định bằng cách so sánh phổ tín hiệu với ngưỡng.
6. Bộ lọc được làm mịn bằng bộ lọc theo tần suất và thời gian.
7. Bộ lọc được áp dụng cho biểu đồ phổ của tín hiệu và được đảo ngược nếu tín hiệu nhiễu không được cung cấp, thuật toán sẽ coi tín hiệu là clip nhiễu, có xu hướng hoạt động khá tốt.

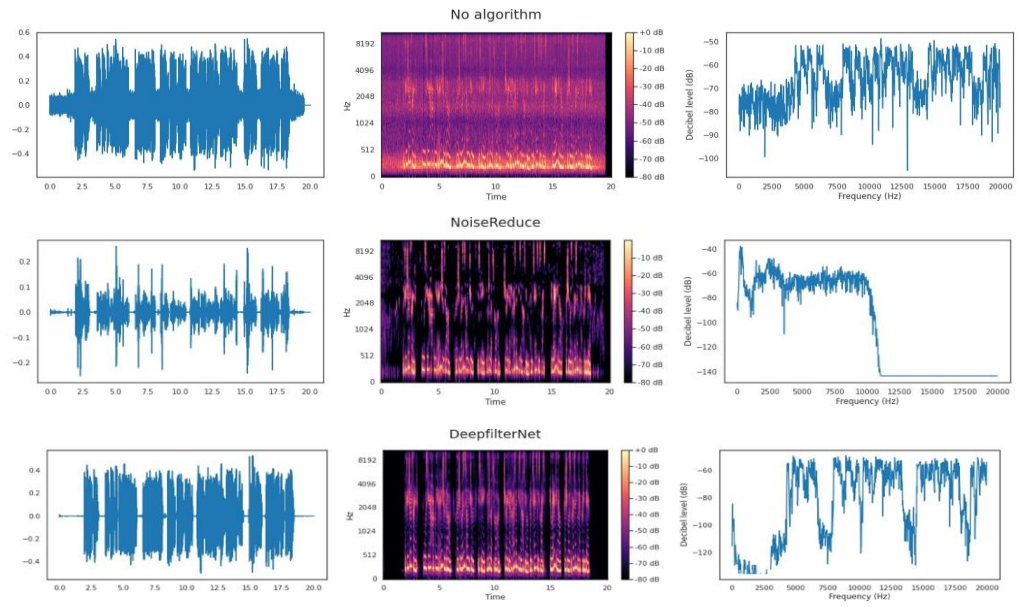
Giảm tiếng ồn không cố định (Non-stationary Noise Reduction). Thuật toán giảm nhiễu không cố định là một phần mở rộng của thuật toán giảm nhiễu cố định, nhưng cho phép cổng nhiễu thay đổi theo thời gian. Thuật toán này được thúc đẩy bởi một phương pháp gần đây trong âm thanh sinh học có tên là bình thường hóa năng lượng trên mỗi kênh. Các bước của thuật toán Giảm tiếng ồn không cố định gồm:

1. Tính toán quang phổ trên tín hiệu.
2. Một phiên quang phổ được làm mịn theo thời gian được tính toán bằng cách sử dụng bộ lọc.
3. Bộ lọc được tính toán dựa trên phổ được làm mịn theo thời gian đó.
4. Bộ lọc được làm mịn bằng bộ lọc theo tần suất và thời gian.
5. Bộ lọc được gắn vào biểu đồ phổ của tín hiệu và được đảo ngược.

a)

b)

c)



Hình 2. Tín hiệu âm thanh trước (a) và sau khi loại bỏ tiếng ồn với NoiseReduce (b) và DeepFilterNet (c)

DeepFilterNet [17] là bộ lọc gồm hai giai đoạn khung sử dụng tính năng "Lọc sâu". Trong giai đoạn đầu tiên, DeepFilterNet tăng cường đường bao quang phổ bằng cách sử dụng mô hình Băng thông hình chữ nhật tương đương

(ERB) mô phỏng nhận thức về tần số của con người. Ngân hàng bộ lọc ERB giảm kích thước đầu vào và đầu ra xuống chỉ còn 32 dải tần cho phép mã hóa/giải mã nhanh trong mạng nội bộ. Tuy nhiên, do băng thông tối thiểu thu được từ 100 Hz đến 250 Hz (tùy thuộc vào kích thước Biến đổi Fourier nhanh (FFT)) thường không đủ để tăng cường các thành phần tuần hoàn, nên việc tăng cường giai đoạn thứ hai dựa vào lọc sâu. Trong giai đoạn thứ hai, lọc sâu được sử dụng để tăng cường các thành phần định kỳ của âm thanh. Quá trình mã hóa được sắp xếp hợp lý trong DeepFilterNet2 chỉ với các lớp tuyến tính được nhóm trên trục tần số thông qua phép nhân ma trận đơn.

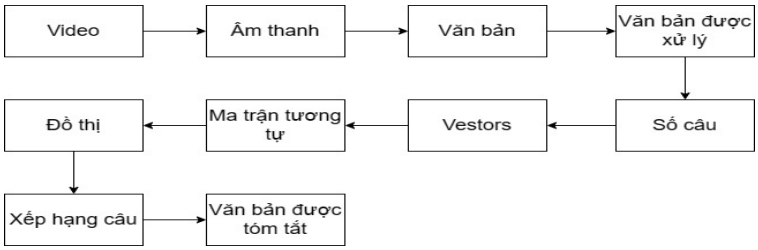
Hình 2a thể hiện dạng sóng âm thanh trước khi sử dụng bất kỳ thuật toán giảm tiếng ồn nào. Sóng nhiễu trông rất nhiều và ảnh hưởng đến chất lượng âm thanh. Quan sát thêm trong Hình 2b kết quả cho thấy số lượng sóng giao thoa đã giảm đáng kể và phương pháp này giảm nhiễu đáng kể. Tuy nhiên, chất lượng âm thanh cũng bị giảm xuống hình ảnh 2b. Chúng ta có thể dễ dàng nhận thấy âm lượng của âm thanh cũng bị giảm đi đáng kể. Quá trình này sẽ làm cho phụ đề video biến mất. Hình 2c hiển thị sóng âm thanh sau khi áp dụng phương pháp deepfilterNet. Kết quả cho thấy phương pháp này đã loại bỏ gần hết nhiễu, chỉ để lại một số điểm sóng tím nhưng với lượng tối thiểu và cũng dễ dàng nhận thấy âm lượng không bị suy giảm.

B. Kỹ thuật phân loại văn bản

Nghiên cứu đã sử dụng Underthesea [20], một bộ công cụ Natural Language Processing (NLP) dành cho người Việt, được phát triển bởi cộng đồng mã nguồn mở. Underthesea có khả năng phân loại văn bản thành chủ đề khác nhau. Bộ công cụ NLP này được đào tạo chuyên biệt và tối ưu hóa cho ngôn ngữ tiếng Việt, giúp đảm bảo phân loại văn bản chính xác và hiệu quả.

C. Các kỹ thuật tóm tắt văn bản

Phương pháp TextRank: TextRank [11] là cách tiếp cận trích xuất và mô tả văn bản không giám sát. Nó là một biến thể của thuật toán PageRank, thường được sử dụng để phân tách từ khóa và tóm tắt nội dung. Trong TextRank, các câu trong văn bản được biểu diễn dưới dạng các đỉnh trong biểu đồ và các cạnh giữa các câu biểu thị mức độ giống nhau giữa chúng, được mô tả ở Hình 3.



Hình 3. Các bước liên quan đến phương pháp tóm tắt Textrank

TextRank xây dựng một biểu đồ từ các từ và mối quan hệ của chúng trong văn bản, sau đó chọn những từ quan trọng nhất dựa trên điểm quan trọng được tính từ toàn bộ biểu đồ. Các bước trong quy trình TextRank như sau:

1. Tiền xử lý văn bản: Văn bản gốc được xử lý bằng cách chuẩn hóa, thêm dấu câu, chính tả và dấu chấm câu.
2. Phân tích câu, từ: Văn bản sau khi tiền xử lý được phân tích thành tập từ cần đánh giá. Các từ được gắn nhãn với loại từ (danh từ, động từ, v.v.) để phân biệt cú pháp được sử dụng.
3. Xây dựng biểu đồ: Mỗi từ được thêm vào biểu đồ và cũng được thêm vào thanh trượt xung quanh từ đó.
4. Xác định mối quan hệ: Mối quan hệ giữa các từ được xác định. Điểm của tất cả các từ được cập nhật dựa trên điểm tương ứng của chúng cho đến khi điểm ổn định. Thuật toán đánh giá được áp dụng trên mỗi đỉnh cho một số lần lặp, thường là từ 20-30 lần lặp.
5. Sắp xếp và lựa chọn từ: Các từ được sắp xếp theo điểm số và chỉ những từ quan trọng nhất được giữ lại (thường là một phần ba số từ).
6. Xử lý hậu kỳ: Cuối cùng, văn bản được xử lý lại bằng cách xem lại danh sách từ gốc và kết hợp các thuật ngữ xuất hiện cùng nhau để tạo thành các cụm từ trong kết quả cuối cùng.

Phương pháp LexRank [10] là một phương pháp không giám sát dựa trên biểu đồ tương tự như TextRank. Nó sử dụng Cosine để tính toán độ tương tự giữa các thuật ngữ. Trọng số của các cạnh trong đồ thị được xác định bởi sự giống nhau giữa hai câu. Phương pháp này bắt đầu bằng cách tạo một biểu đồ với các đỉnh biểu thị các câu trong văn bản và các cạnh biểu thị mối quan hệ chặt chẽ giữa các câu. Mức độ giống nhau giữa các câu được đo bằng mô hình túi từ, trong đó tần suất xuất hiện của các từ sẽ xác định mức độ giống nhau. Term Frequency-Inverse Document Frequency (TF-IDF) được sử dụng để đo tần suất từ, trong đó TF đóng góp vào sự tương tự khi số lần xuất hiện của từ đó lớn hơn. Tương quan nghịch đảo của các từ trong văn bản cũng được tính toán. Phương pháp thực hiện đo khoảng cách giữa hai câu x và y. Các khoảng cách của các cặp câu này tạo thành ma trận tương tự, có thể được sử dụng xây dựng đồ thị. LexRank xác định tầm quan trọng của các câu dựa trên tầm quan trọng của chúng đối với các câu liền kề. Để chọn những câu quan trọng nhất, phương pháp sử dụng một ngưỡng từ ma trận tương tự. Mối quan hệ giữa các cặp câu có trọng số nhỏ hơn ngưỡng sẽ bị loại bỏ. Kết quả là tập con của đồ thị độ tương tự, từ đó chọn câu có số bậc lớn nhất là quan trọng và có thể dùng làm câu tóm tắt của văn bản.

Phương pháp Latent Semantic Analysis (LSA) [9] là một kỹ thuật không giám sát để tạo ra các tóm tắt văn bản ngữ nghĩa từ một tập hợp các từ là phân tích ngữ nghĩa tiềm ẩn. LSA hoạt động bằng cách cô đặc kích thước dữ liệu thành một khu vực nhỏ hơn trong khi vẫn giữ lại dữ liệu quan trọng.

Phương pháp Luhn. Một cách quan trọng để tạo tóm tắt từ một tập hợp các từ là phương pháp Luhn [7]. Phương pháp này có thể được thực hiện theo hai bước. Trong bước đầu tiên, phân tích tần suất để tìm các thuật ngữ quan trọng đối với ý nghĩa của tài liệu và tìm kiếm các từ có ý nghĩa nhưng không phổ biến trong tiếng Anh. Trong giai đoạn thứ hai, xác định nhóm thuật ngữ cơ bản nhất của tài liệu và chọn một tập hợp con loại trừ các từ phổ biến nhưng quan trọng. Ba hành động có thể được thực hiện để thực hiện điều này:

1. Chuyển phần thân của câu thành biểu diễn số hoặc vectơ bằng cách tránh các từ thông thường.
2. Đếm số từ quan trọng còn lại, loại bỏ các từ như "và" và dừng các từ bằng phương pháp đánh giá câu.
3. Xác định xếp hạng cho các câu và chọn câu có xếp hạng cao nhất làm tóm tắt.

Phương pháp Edmundson [8] là một phần mở rộng của phương pháp Luhn kết hợp công việc thống kê trước đó. Edmundson đã bổ sung thêm ba phương pháp để đo lường tầm quan trọng của các câu. Phương pháp này phụ thuộc nhiều vào ngôn ngữ và yêu cầu một danh sách các "từ tốt" và " từ xấu" để loại trừ các từ dừng. Những từ tốt, còn được gọi là từ thực tế, có ý nghĩa tích cực và bao gồm so sánh, châm ngôn, trạng từ kết thúc, thuật ngữ giá trị, câu hỏi tương đối và các từ nhân quả, chẳng hạn như "quan trọng", "tuyệt vời", "nổi tiếng", "vinh quang", "thú vị" ," hoặc "tuyệt vời". Những từ xấu, có ý nghĩa tiêu cực, bao gồm các cách diễn đạt phản xạ, cách diễn đạt châm biếm, cách diễn đạt tầm thường và cách diễn đạt mơ hồ, chẳng hạn như "khó khăn" hoặc "không thể". Các từ trống không liên quan và bao gồm từ thứ tự, số thứ tự, động từ, giới từ, đại từ, tính từ, trợ động từ, mạo từ và liên từ,...

Phương pháp Kullback-Lieber Sum (gọi tắt KL) [21]. Phương pháp thêm các câu vào bản tóm tắt theo cách tiếp cận "háu ăn" (heuristic). Mục tiêu là tìm ra bản tóm tắt của văn bản bao gồm một bộ câu có độ dài ngắn hơn L từ và phân phối *unigram (n-gram)* của bộ từ đó là gần giống với văn bản gốc.

IV. KẾT QUẢ THỰC NGHIỆM

Bộ dữ liệu tin tức để thực nghiệm bao gồm hơn 100 video tin tức tổng hợp tin tức hàng ngày từ Youtube. Bộ dữ liệu gồm tiếng Việt và tiếng Anh với 9 chủ đề khác nhau. Bảng 1 thể hiện thông tin về bộ dữ liệu được thực nghiệm cho thấy được sự đa dạng về chủ đề, độ dài, số chữ trong các video và các thông số liên quan khác như số chữ và số dòng sau khi tự động tóm tắt thì chúng ta có thể thấy được sự ngắn gọn mà đoạn văn bản tóm tắt mang lại.

Bảng 1. Thông tin về dữ liệu thực nghiệm và số chữ sau khi thực hiện tóm tắt

Chủ đề	Số video	Tổng thời lượng video (giây)	Trung bình số CHỮ trong video (gốc)	Trung bình số CÂU trong video (gốc)	Trung bình số CHỮ sau khi tóm tắt	Trung bình số CÂU sau khi tóm tắt
Chính trị	14	2382,49	439	12	192	3
Đời sống	10	1806,91	395	19	183	5
Khoa học	10	2859,75	682	35	324	8
Kinh doanh	2	313,37	398	20	210	5
Pháp luật	16	1287,30	239	7	113	2
Sức khỏe	5	268,62	172	7	76	1
Thế giới	30	5508,71	549	28	259	7
Thể thao	16	977,40	228	7	86	2
Vi tính	20	3013,91	51	3	21	1

Đối với kịch bản 1, chúng tôi đo lường mức độ tương tự giữa văn bản gốc và văn bản được phát hiện trong video (là kết quả dự đoán). Chúng tôi tính toán mức độ giống nhau của văn bản được phát hiện từ tập tin âm thanh và so sánh văn bản gốc với độ đo cosine, với các kết quả độ tương tự trong phạm vi [0,1].

Đối với các thực nghiệm trên các bài toán tóm tắt văn bản trong kịch bản 2 và kịch bản 3 chúng tôi sử dụng các độ đo Precision, Recall và F1-score với các cách tiếp cận ROUGE-1, ROUGE-2 và ROUGE-L để đánh giá hiệu suất.

A. Kịch bản 1: So sánh hiệu suất của các kỹ thuật loại bỏ tiếng ồn

Bảng 2. Kết quả so sánh các thuật toán loại bỏ tiếng ồn trên 1 video tin tức được thực nghiệm

Thuật toán	Độ dài văn bản gốc	Độ dài được phát hiện	Điểm tương đồng
Không loại bỏ tiếng ồn	283	142	0,7266
NoiseReduce	283	171	0,7339
DeepFilterNet	283	184	0,8442

Trong Bảng 2 thể hiện rằng các kỹ thuật giảm tiếng ồn như NoiseReduce hoặc DeepFilterNet có thể cải thiện đáng kể độ chính xác của chuyển đổi lời nói thành văn bản trong môi trường ồn ào với độ dài trong tất cả các bảng

được đo bằng số lượng từ. Các kết quả cho thấy tầm quan trọng của việc giảm tiếng ồn. Tiếng ồn có thể ảnh hưởng đến việc chuyển lời nói trong video thành văn bản. Bên cạnh đó, ta cũng thấy hiệu quả của các phương pháp giảm tiếng ồn khác nhau.

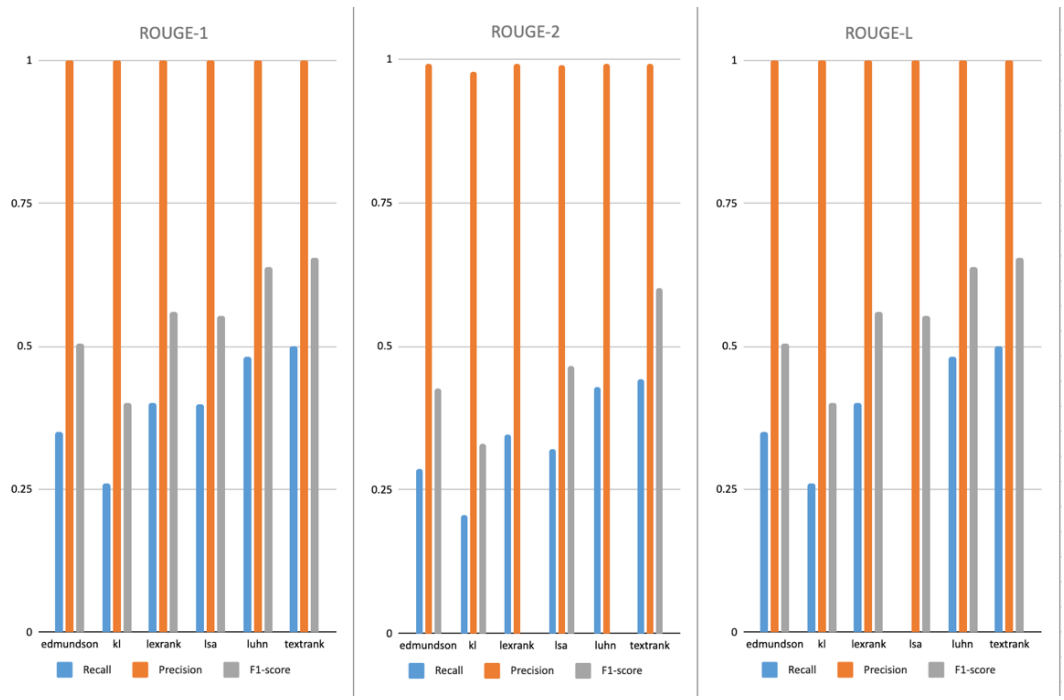
B. Kịch bản 2: So sánh các phương pháp tóm tắt văn bản

Dựa vào bảng Bảng 3, ta có thể rút ra các nhận xét về hiệu suất của các kỹ thuật tóm tắt văn bản. Đầu tiên, tất cả các kỹ thuật nêu trên đều có độ chính xác (độ đo precision) là (1.00) Điều này cho thấy các kỹ thuật này đều có khả năng tạo ra các câu tóm tắt chính xác và liên quan đến nội dung ban đầu. Nhận xét từ kết quả, ta thấy, Phương pháp Edmundson cho thấy sự cải thiện với mức điểm F1 0,504 trên ROUGE-1 và 0,428 trên ROUGE-2. Thêm nữa, Phương pháp KL và LSA đạt kết quả tương đối tốt, với điểm F1 dao động từ 0,400 đến 0,553 đối với ROUGE-1 và từ 0,207 đến 0,321 đối với ROUGE-2. Với LexRank và TextRank, chúng tiếp tục thể hiện hiệu suất ấn tượng, với điểm F1 dao động từ 0,559 đến 0,655 trên ROUGE-1 và từ 0,346 đến 0,444 trên ROUGE-2. Ta thấy rằng, Phương pháp Luhn vẫn là phương pháp hoạt động tốt nhất với điểm số F1 cao nhất trên ROUGE-1 và ROUGE-2 lần lượt là 0,638 và 0,428.

Bảng 3. Kết quả độ chính xác với 6 phương pháp tóm tắt văn bản

CÁC KỸ THUẬT TÓM TẮT VĂN BẢN	ROUGE-1			ROUGE-2			ROUGE-L		
	Recall	Precision	F1-score	Recall	Precision	F1-score	Recall	Precision	F1-score
Edmundson	0,350	1,000	0,504	0,286	0,992	0,428	0,350	1,000	0,504
KL	0,259	1,000	0,400	0,207	0,978	0,329	0,259	1,000	0,400
Lexrank	0,402	1,000	0,559	0,346	0,991	0,497	0,402	1,000	0,559
LSA	0,399	1,000	0,553	0,321	0,989	0,466	0,398	1,000	0,553
LUHN	0,482	1,000	0,638	0,428	0,992	0,585	0,482	1,000	0,638
Textrank	0,500	1,000	0,655	0,444	0,992	0,601	0,500	1,000	0,655

Tuy nhiên, khi xem xét các chỉ số recall và F1-score, ta thấy rằng các kỹ thuật có sự khác biệt trong việc tái tạo lại toàn bộ nội dung ban đầu. Cụ thể, Edmundson và KL có recall thấp hơn so với các kỹ thuật khác, điều này cho thấy chúng không tái tạo lại được một phần lớn nội dung của văn bản gốc. Trong khi đó, Lexrank, LSA, LUHN và Textrank đều có recall cao, cho thấy khả năng tái tạo lại nội dung ban đầu tốt hơn. Tuy nhiên, Lexrank và Textrank có F1-score tương đối cao hơn so với LSA và LUHN, cho thấy chúng có sự cân đối giữa khả năng tái tạo lại nội dung và độ chính xác. Tóm lại, dựa vào bảng trên, Lexrank và Textrank được coi là hai kỹ thuật tóm tắt văn bản có hiệu suất tốt nhất với sự cân đối giữa khả năng tái tạo lại nội dung và độ chính xác cao cụ thể chúng ta có thể thấy được kết quả như sau:



Hình 4. So sánh độ chính xác, Recall, Precision, F1-score của 6 phương pháp tóm tắt văn bản với 3 cách đánh giá (ROUGE-1, ROUGE-2, ROUGE-L)

C. Kịch bản 3: So sánh kết quả tóm tắt văn bản của các chủ đề video

Phân tích chi tiết kết quả độ chính xác cho từng chủ đề trong Bảng 4 cho thấy mức độ tóm tắt văn bản đạt được ở chủ đề "Chính trị" và "Khoa học" là rất cao. Cả hai chủ đề này đều thể hiện các chỉ số ROUGE-1, ROUGE-2 và ROUGE-L ở mức vượt qua ngưỡng 0,5, đánh dấu sự thành công trong việc giữ lại cả nội dung cũng như cấu trúc quan trọng của văn bản gốc. Đáng chú ý, các chỉ số ROUGE-2 và ROUGE-L còn đạt kết quả gần 0,5, thể hiện sự tương đồng cao về ngôn ngữ và cấu trúc giữa văn bản tóm tắt và văn bản gốc.

Bảng 4. Kết quả độ chính xác với 9 chủ đề khác nhau dùng cho tóm tắt văn bản

Chủ đề	ROUGE-1			ROUGE-2			ROUGE-L		
	Recall	Precision	F1-score	Recall	Precision	F1-score	Recall	Precision	F1-score
Chính trị	0,410	1,000	0,560	0,358	0,993	0,504	0,410	1,000	0,563
Đời sống	0,391	1,000	0,550	0,327	0,985	0,479	0,391	1,000	0,550
Khoa học	0,405	1,000	0,563	0,325	0,978	0,470	0,405	1,000	0,561
Kinh doanh	0,389	1,000	0,537	0,315	0,991	0,453	0,389	1,000	0,537
Pháp luật	0,403	1,000	0,555	0,354	0,990	0,504	0,403	1,000	0,555
Sức khỏe	0,366	1,000	0,507	0,318	0,996	0,449	0,366	1,000	0,507
Thế giới	0,404	1,000	0,562	0,330	0,982	0,486	0,404	1,000	0,562
Thể thao	0,398	1,000	0,540	0,350	0,994	0,483	0,398	1,000	0,540
Vi tính	0,294	1,000	0,443	0,244	1,000	0,382	0,294	1,000	0,443

V. KẾT LUẬN

Trong nghiên cứu này, chúng tôi đã thực hiện đánh giá các phương pháp để tóm tắt nội dung từ video trên dữ liệu được tập hợp từ 9 chủ đề các bản tin tức. Các thuật toán giảm tiếng ồn như Noisereduce và DeepFilterNet, bộ công cụ Underthesea, cùng các kỹ thuật để tóm tắt văn bản đã được sử dụng để phân tích so sánh kết quả. Chúng tôi đã đánh giá kết quả qua 3 cách tiếp cận Rouge-1, Rouge-2, và Rouge-L. Qua các độ đo, cho thấy các bản tin về thế giới và khoa học cho độ chính xác tóm tắt cao nhất và Textrank cho kết quả tốt nhất trong 6 phương pháp tóm tắt được khảo sát.

Hướng phát triển sắp tới, các nghiên cứu có thể tăng bộ dữ liệu thực nghiệm có thể bao gồm nhiều video có nhiều yếu tố nhiễu hơn để tăng cường kỹ thuật xử lý tiếng ồn cũng như bổ sung thêm chủ đề để việc phân loại đa dạng hơn. Đồng thời, chúng ta có thể phát triển kết quả Phân loại chủ đề và xử lý tóm tắt văn bản trong video để tạo ra phụ đề cho các video trong ngành phim ảnh, phát thanh truyền hình và các sản phẩm lưu trữ dữ liệu, phục vụ các đối tượng khiếm thính, nghiên cứu dữ liệu. Kết quả hiện tại vẫn chưa cao chủ yếu do độ chính xác chuyển giọng nói thành văn bản có thể chưa cao và còn có thể gây nhiễu do tiếng ồn. Tuy nhiên, kết quả tóm tắt văn bản video được kỳ vọng có thể cải thiện trong tương lai để dùng làm cơ sở để phát triển tóm lược video phục vụ quảng bá, lưu trữ và truy xuất, tìm kiếm dữ liệu.

TÀI LIỆU THAM KHẢO

[1] U. Shrawankar and V. Thakare, "Noise Estimation and Noise Removal Techniques for Speech Recognition in Adverse Environment," in *Intelligent Information Processing V*, Springer Berlin Heidelberg, 2010, pp. 336-342. DOI: 10.1007/978-3-642-16327-2_40.

[2] H. T. Nguyen, T. N. L. Thanh, T. L. Ngoc, A. D. Le, and D. T. Tran, "Evaluation on Noise Reduction in Subtitle Generator for Videos," in *Innovative Mobile and Internet Services in Ubiquitous Computing*, Springer International Publishing, 2022, pp. 140-150. DOI: 10.1007/978-3-031-08819-3_14.

[3] P. Singh, P. Chhikara, and J. Singh, "An Ensemble Approach for Extractive Text Summarization," in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, 2020, pp. 1-7. DOI: 10.1109/ic-ETITE47903.2020.95.

[4] M. N. Azadani, N. Ghadiri, and E. Davoodijam, "Graph-based biomedical text summarization: An itemset mining and sentence clustering approach," *J. Biomed. Inform.*, vol. 84, pp. 42-58, Aug. 2018, DOI: 10.1016/j.jbi.2018.06.005.

[5] J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," in *2019 International Conference on Data Science and Communication (IconDSC)*, 2019, pp. 1-3. DOI: 10.1109/IconDSC.2019.8817040.

[6] W. Waseemullah *et al.*, "A Novel Approach for Semantic Extractive Text Summarization," *Appl. Sci.*, vol. 12, no. 9, p. 4479, Apr. 2022, DOI: 10.3390/app12094479.

[7] H. P. Luhn, "The Automatic Creation of Literature Abstracts," *IBM J. Res. Dev.*, vol. 2, no. 2, pp. 159-165, 1958, DOI: 10.1147/rd.22.0159.

[8] H. P. Edmundson, "New Methods in Automatic Extracting," *J. ACM*, vol. 16, no. 2, pp. 264-285, Apr. 1969, DOI: 10.1145/321510.321519.

[9] M. G. Ozsoy, F. N. Alpaslan, and I. Cicekli, "Text summarization using Latent Semantic Analysis," *J. Inf. Sci.*, vol. 37, no. 4, pp. 405-417, Jun. 2011, DOI: 10.1177/0165551511408848.

[10] G. Erkan and D. R. Radev, "LexRank: Graph-Based Lexical Centrality as Salience in Text Summarization," *J Artif Int Res*, vol. 22, no. 1, pp. 457-479, Dec. 2004.

[11] R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Text," in *Conference on Empirical Methods in Natural Language Processing*, 2004. [Online]. Available: https://api.semanticscholar.org/CorpusID:577937

[12] S. Dalal, A. Singhal, and B. Lall, "LexRank and PEGASUS Transformer for Summarization of Legal Documents," in *Lecture Notes in Electrical Engineering*, Springer Nature Singapore, 2023, pp. 569-577. DOI: 10.1007/978-981-99-0085-5_46.

[13] K. Ramani, K. Bhavana, A. Akshaya, K. S. Harshita, C. R. Thoran Kumar, and M. Srikanth, "An Explorative Study on Extractive Text Summarization through k-means, LSA, and TextRank," in *2023 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, 2023, pp. 1-6. DOI: 10.1109/WiSPNET57748.2023.10134303.

[14] S. Tomar, "Converting video formats with FFmpeg," *Linux J.*, vol. 2006, no. 146, p. 10, 2006.

[15] T. Sainburg, M. Thielk, and T. Q. Gentner, "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires," *PLOS Comput. Biol.*, vol. 16, no. 10, p. e1008228, Oct. 2020, DOI: 10.1371/journal.pcbi.1008228.

[16] T. Sainburg, "timsainb/noisereduce: v1.0." Zenodo, Jun. 2019. DOI: 10.5281/zenodo.3243139.

[17] H. Schröter, T. Rosenkranz, A. N. Escalante-B., and A. Maier, "Deep Multi-Frame Filtering for Hearing Aids," in *INTERSPEECH*, 2023.

[18] A. Zhang (Uberi), "SpeechRecognition: Library for performing speech recognition, with support for several engines and APIs, online and offline." Accessed: Aug. 09, 2023. [MacOS :: MacOS X, Microsoft :: Windows, Other OS, POSIX :: Linux]. Available: https://github.com/Uberi/speech_recognition#readme

[19] O. Tilk and T. Alumäe, "Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration," in *Interspeech 2016*, ISCA, Sep. 2016. DOI: 10.21437/interspeech.2016-1517.

[20] Anh Vu and Nguyen Dang Duc Tai and Bui Nhat Anh and Vuong Quoc Binh and Doan Viet Dung, "Underthesea Documentation," 2018. {https://underthesea.readthedocs.io/en/latest/index.html}

[21] J. M. Joyce, "Kullback-Leibler Divergence," in *International Encyclopedia of Statistical Science*, Springer Berlin Heidelberg, 2011, pp. 720-722. DOI: 10.1007/978-3-642-04898-2_327.

AN APPROACH FOR SUMMARIZATION OF VIDEO NEWSLETTERS

Hai Thanh Nguyen, Khoa Viet Le, Toan Khanh Do, Nguyen Thai Nghe

ABSTRACT: *With the explosion of the Internet, sources of information are increasingly numerous and diverse, from which the need to classify and summarize information becomes increasingly necessary. Newsletters from Videos are a precious source of information to help us stay updated with the latest news. However, with relatively long videos, while life in urban areas is busy, it is difficult for users to watch the entire video. In addition, the need to seek content from video is also essential. Summary of video content can help users quickly grasp video topics and content and support search engines. In this study, we evaluated the effectiveness of 6 text summarization methods on videos gathered from 9 popular news topics. Before summarizing, we have filtered noise to get better-extracted text content. Experimental results show that Textrank gives the best results. We see that noise reduction in video newslettlers. In addition, the summarization performances on video related to international news and Science reveal the best. The proposed method is expected to continue to be improved in content summary processing applied to broadcasters and daily newsletter.*