Logo Image not found or type unknown

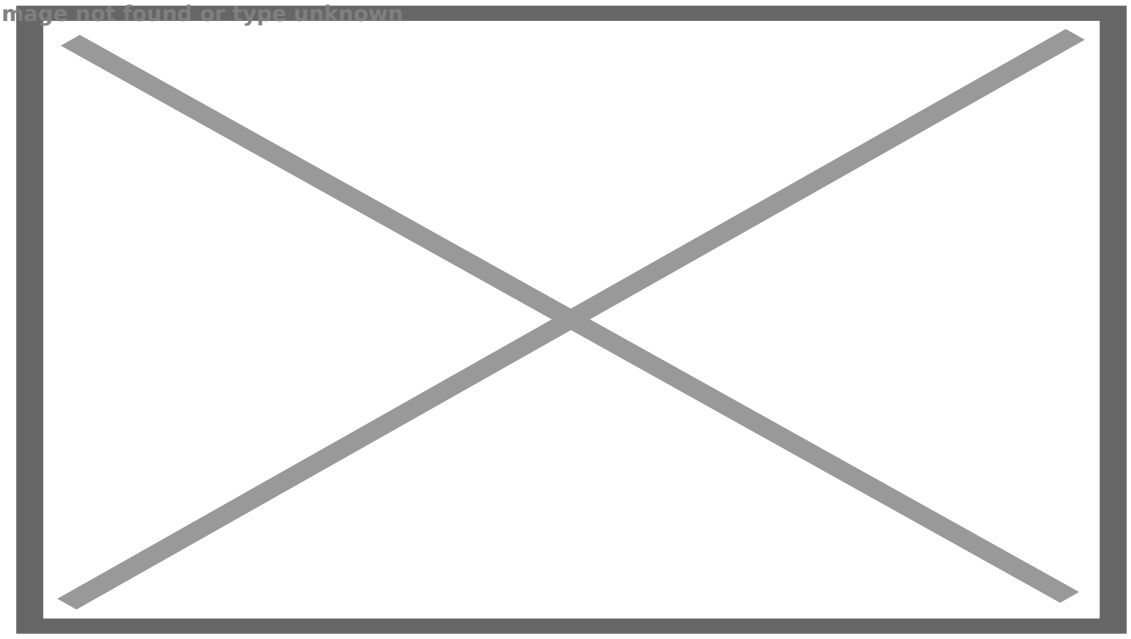
Báo chí dữ liệu và công nghệ tự động làm báo

18:09 15/11/2018

Tác giả: Đang cập nhật

Beritagar.id là một tờ báo điện tử bằng tiếng Bahasa, có trụ sở chính tại Thủ đô Jakarta của Indonesia. Đây là tờ báo tiên phong của khu vực ASEAN trong việc sử dụng những công nghệ hiện đại.

Image not found or type unknown



Trang Báo chí rô-bốt của Báo điện tử Beritagar.id

Báo điện tử Beritagar.id ra đời ngày 24/8/2015, trên cơ sở hợp nhất trang báo điện tử cũ Beritagar.com và công ty dữ liệu LokaData. Sau 3 năm hoạt động, đến nay Beritagar.id có 100 nhân viên, gồm có 25 nhà báo, 25 lập trình viên, và 50 người làm kinh doanh, tiếp thị, hành chính, kỹ thuật.

“Khi chúng tôi mới thành lập năm 2013, thị trường báo điện tử ở Indonesia đã bão hòa. Chúng tôi quyết định phải tìm một lối đi riêng, không lặp lại các mô hình báo điện tử đã có. Chính vì vậy, chúng tôi chọn báo chí dữ liệu và áp dụng triệt để công nghệ tự động trong làm báo” - Ông Rahadian Paramita, Phó Tổng Biên tập Báo điện tử Beritagar chia sẻ.

Máy tính tự viết báo (Robotaria)

Beritagar có hai công đoạn quan trọng nhất để máy tính tự động xuất bản một bài báo, đó là: tự động cập nhật nguồn số liệu lớn; trình bày số liệu dưới dạng đồ họa và chữ viết.

Nguồn số liệu:

Phó Tổng Biên tập Báo điện tử Beritagar cho biết: “Chúng tôi mua số liệu của các công ty chuyên cung cấp dữ liệu lớn. Từ đó, tòa soạn có thông tin cùng lúc với các cơ quan sở hữu thông tin đó. Ví dụ: các kết quả thi đấu thể thao, số liệu quan trắc thời tiết, biến động trên sàn chứng khoán, v.v..”.

“Ở Beritagar, chúng tôi có một nguyên tắc là không dùng dữ liệu thứ cấp, ví dụ số liệu từ các bản báo cáo, hay các cuộc điều tra khảo sát”. Ông Rahadian cho rằng, con số từ các cuộc khảo sát không hoàn toàn chính xác, và nó thể hiện quan điểm của người cung cấp kết quả khảo sát. Quan điểm đó có thể cảm tính hoặc không khách quan. Thậm chí kết quả khảo sát trước bầu cử có thể bị nhào nặn theo ý đồ của người tranh cử”. Chính vì vậy, dữ liệu đầu vào phải ở dạng còn thô (raw data), chưa bị can thiệp.

Biến dữ liệu thành nội dung:

Ngay khi máy tính của Beritagar nhận dữ liệu, phần mềm sẽ tự động lọc ra các con số có ý nghĩa để điền vào những mẫu câu có sẵn, tự đổ số liệu vào các cột biểu bảng, sau đó xuất ra hình đồ thị và nội dung bài, và tự động đưa bài viết lên trang web.

Dưới những bài viết tự động, luôn có dòng chữ tuyên bố: “Đây là bài báo do chương trình máy tính chuyển số liệu thành chữ viết và đồ thị và tự động xuất bản”.
Khi những bài báo này được dẫn link để đăng trên các trang truyền thông xã hội của Beritagar, cuối phần giới thiệu bài báo luôn đi kèm hashtag #Robotaria (Báo chí rô bốt) để phân biệt với những bài báo do phóng viên viết.

Thời gian từ khi sự kiện xảy ra, đến lúc máy đo được dữ liệu và xử lý để xuất bản, kéo dài không quá 5 phút. Với những trận thi đấu, thời gian từ khi có diễn biến trận đấu đến khi diễn biến đó được xuất hiện trên trang báo điện tử chỉ còn tính bằng giây.

Để rút ngắn thời gian xuất bản tự động, các phóng viên của Beritagar đã mất một năm để dạy máy tính tự làm báo. Hàng trăm ngàn bài báo đã được đưa cho máy đọc để tự tìm ra quy tắc về mẫu câu và cấu trúc bài. Hàng ngàn mẫu câu đã được viết sẵn để máy tự điền kết quả.

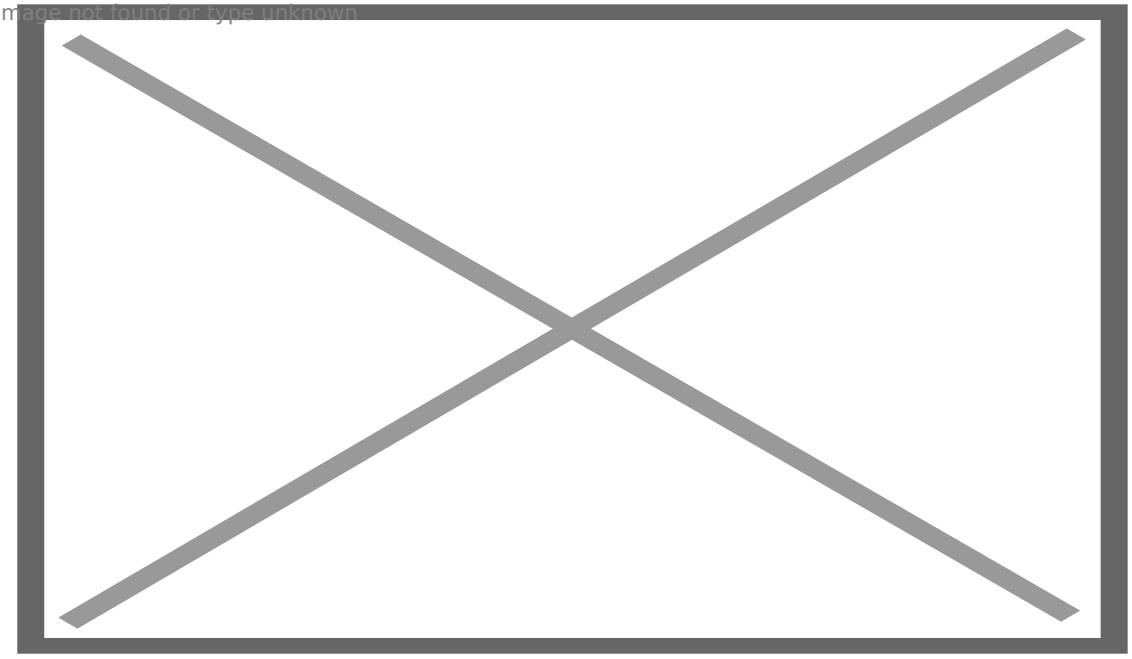
Ví dụ, nếu tỉ số trận đấu là 0 - 1 thì phần thắng thuộc về đội nào? Kết quả 0 - 1 là thắng, nhưng kết quả cách biệt 0 - 5 thì máy cần điền tỷ số vào mẫu câu có sẵn cụm từ thắng đậm, hoặc thua đau. Beritagar quy định bài báo cần viết theo mô hình kim cương. Sau câu (hoặc đoạn) đầu tiên nêu chủ đề của bài báo, câu (hoặc đoạn) thứ 2 cần cho bạn đọc biết thông tin quan trọng nhất, là góc độ đưa tin. Thông tin trong những câu (hoặc đoạn) sau đó được xếp theo mức độ quan trọng giảm dần. Quy tắc này áp dụng cho cả bài viết của phóng viên và bài do máy tính tự viết. Tòa soạn cũng chọn những bài báo có cấu trúc mô hình kim cương để dạy cho máy học.

Ông Rahadian Paramita cho biết 80% số bài báo đưa cho máy tính đọc là những bài báo cũ, 20% còn lại là những bài báo mới, lạ, không theo cấu trúc thông thường. Mục đích là để máy không lặp lại theo lối mòn và máy tự rèn khả năng sáng tạo. Beritagar có các bản tin ngày và bản tin tuần. Những tin tức thời sự do máy tính tự động xuất bản thường là các tin hằng ngày. Với những tin tuần, nhà báo có thời gian để đầu tư điều tra sâu hơn, phỏng vấn nhân vật và kiểm chứng những ý kiến do nhân vật đưa ra.

Ở giai đoạn phát hiện đề tài, Beritagar cung cấp dữ liệu phân tích nội dung báo chí. Chương trình riêng của báo tự động thu thập và phân tích 1.000 bài báo mới xuất bản trên các trang báo điện tử. Sau đó, máy tính tìm ra những chủ đề đang được đề cập nhiều nhất và các góc độ đưa tin. Chương trình phần mềm phân tích nội dung báo chí chính là bộ não của tòa soạn.

Sau 3 năm hoạt động, bộ não này có khả năng phân tích đúng từ 86% đến 96% so với con người, tuỳ vào từng chủ đề báo chí. Dựa vào những phân tích đó, phóng viên quyết định sẽ chọn đề tài gì và đưa tin từ góc độ nào, để tránh trùng lặp những góc độ đã được các báo khác khai thác.

Image not found or type unknown



Tòa soạn điện tử Beritagar.id. Ảnh: TL

Yêu cầu đối với nhà báo

Đặc trưng riêng của bài báo trên trang Beritagar là luôn có phần hình ảnh bắt mắt và chuyên nghiệp, có thể là ảnh chụp, phim ngắn, tranh vẽ và đồ thị. Để có sản phẩm đa phương tiện, người làm báo phải làm việc theo nhóm, gồm: phóng viên, kỹ sư lập trình và họa sĩ.

Chỉ cần 10 nghìn USD vốn ban đầu, đã có thể gây dựng cơ sở hạ tầng, sắm thiết bị kỹ thuật, lắp đặt máy chủ và mua số liệu. Khoản đầu tư tốn kém nhất vẫn là đầu vào con người, đào tạo nhân sự để thích hợp với cách làm báo chí dữ liệu.
Trong tuyển dụng, tờ báo vẫn dựa vào những tiêu chí rất cơ bản: sử dụng thành thạo ngôn ngữ Bahasa, có năng khiếu kể chuyện, có khả năng phát hiện đề tài và kiểm chứng thông tin. Trong đó, kỹ năng viết vẫn là điều kiện tiên quyết để trở thành nhà báo của nhánh báo chí dữ liệu.

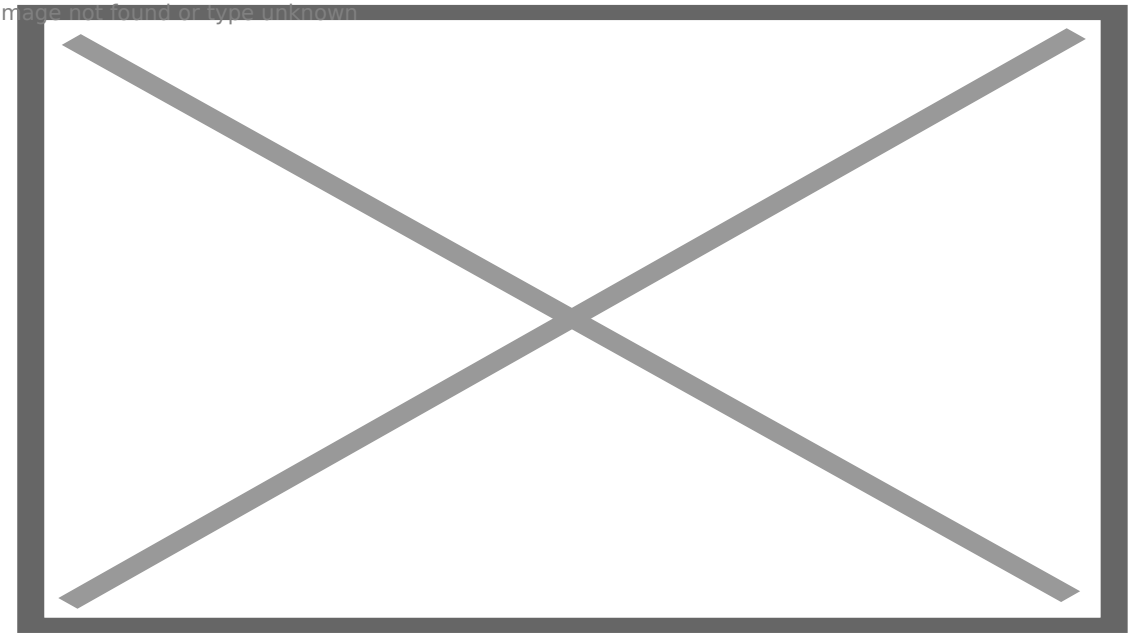
Sau tuyển dụng, Beritagar đào tạo để nhà báo có thể kiêm lập trình (journal-coder). Họ cần có những kỹ năng lập trình máy tính đơn giản để dạy máy viết theo văn phong riêng của mỗi phóng viên, hoặc để phối hợp hiệu quả với những lập trình viên chuyên nghiệp.

Sáng thứ 7 hằng tuần, phóng viên của Beritagar tham dự lớp báo chí dữ liệu được mở ngay tại tòa soạn. Nhà báo Aghnia Adzkia của báo Beritagar là người dạy những lớp này. Cô Aghnia tốt nghiệp thạc sĩ ngành báo điện tử tại viện Goldsmiths, thuộc Đại học London, Anh. Tuy báo chí dữ liệu chỉ là một môn học trong chương trình này, cô Aghnia đã kịp kết nối với những chuyên gia khoa học dữ liệu hàng đầu thế giới. Các khóa tập huấn do Aghnia phụ trách thường xuyên mời chuyên gia

trao đổi qua mạng với phóng viên, dạy cho phóng viên biết đọc số liệu, phát hiện những điểm quan trọng từ dữ liệu thô và biết cách khai thác mỏ dữ liệu lớn.

Nhận ra tầm quan trọng của các kỹ năng phân tích dữ liệu, các trường đại học báo chí của Indonesia đã cử giảng viên đến tòa soạn báo Beritagar để tham gia các khóa tập huấn hằng tuần do Aghnia tổ chức.

Image not found or type unknown



Các bài báo trên trang Beritagar là luôn có phần hình ảnh bắt mắt và chuyên nghiệp. Ảnh chụp màn hình

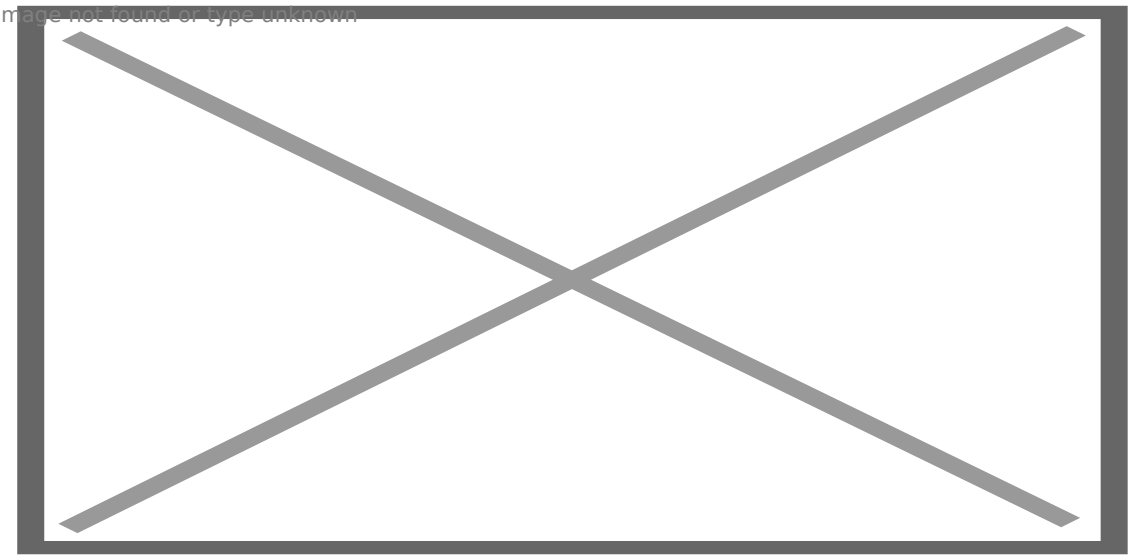
Tòa soạn kiếm tiền như thế nào? Trang báo Beritagar.id hoàn toàn không đăng quảng cáo. Tòa soạn dùng những vị trí hiển thị banner để quảng bá những chương trình, sự kiện của báo và giới thiệu các công ty đối tác.

Ngay từ ban đầu, Beritagar xác định chiến lược phát triển những nội dung nghiêm túc, tập trung chủ đề kinh tế, chính trị, tránh những bài viết giật gân, giải trí, không đăng tin về người nổi tiếng. Chính vì vậy, tuy rất có uy tín trong giới báo chí và công nghệ, Beritagar vẫn không được nhiều người biết đến. Chiến lược này vẫn sẽ được duy trì trong tương lai. Không kiếm tiền từ quảng cáo, và cũng không nhắm tới mục tiêu tăng trưởng về số lượng người truy cập, vậy Beritagar kiếm tiền bằng cách nào?

Như đã trình bày từ phần trước, Beritagar gồm có hai hợp phần là trang báo điện tử và công ty dữ liệu LokaData. Nguồn thu của tờ báo lấy hoàn toàn từ LokaData. LokaData mua dữ liệu lớn ở dạng thô, chưa xử lý. Một phần dữ liệu này được máy tính tự động phân tích để xuất bản bài báo.

Phần dữ liệu còn lại sẽ được các kỹ sư dữ liệu (data analist) và phóng viên viết lại thành những báo cáo, hoặc xây dựng cơ sở dữ liệu. Nhờ vậy, Beritagar có nguồn tài nguyên phong phú gồm những thông tin về nhân khẩu học của thị trường Indonesia rộng lớn với dân số gần 300 triệu người sống tại 500 tỉnh thành. Các doanh nghiệp cần tìm hiểu những thông tin thị trường mới nhất thường tìm mua dữ liệu của Beritagar. Thậm chí, chương trình thành phố thông minh Jakarta, một đề án được chính phủ đầu tư, cũng là khách hàng mua dữ liệu từ Beritagar.

Image not found or type unknown



Tương lai của báo chí dữ liệu

Theo phóng viên Aghnia Adzkia, báo chí dữ liệu không phải là cách làm mới mẻ. Từ thập niên 70, các tờ báo quốc tế và Indonesia đã dùng biểu đồ kèm theo bài viết, nhất là trong những dịp đưa tin về bầu cử. Cùng với sự phát triển của kỹ thuật số và Internet, số liệu lớn đang trở thành một mỏ dồi dào, cần người biết khai thác và tận dụng.

Aghnia cho biết, từ khi có kỹ năng phân tích dữ liệu, hiệu suất công việc của cô tăng lên gấp 3 so với trước đây. Kèm theo thu nhập cũng tăng tương ứng. Tuy nhiên, không phải nhà báo nào cũng kiên trì học và áp dụng các kỹ năng làm việc với những con số, nhất là những nhà báo của những tòa soạn không bắt buộc dùng dữ liệu.

Như vậy, chiến lược phát triển báo chí dữ liệu, xuất bản báo tự động, bán cơ sở dữ liệu và các dịch vụ phân tích dữ liệu sẽ khó trở thành xu hướng đại trà ở nhiều cơ quan báo chí. Nhưng làn sóng sử dụng công nghệ 4.0 đã và đang lan rộng đến mọi ngành nghề, trong đó có nghề báo. Tận dụng dữ liệu lớn sẽ là cách thức lướt sóng cho những tòa soạn định hướng phát triển bằng công nghệ mới. Để dữ liệu khách quan lên tiếng, tránh cảm tính chủ quan, là một cách làm báo tôn trọng tuyệt đối nhu cầu thông tin của độc giả./.

Mạch Lê Thu (NCS Đại học Monash, Úc)

Một số link để tòa soạn tham khảo, lấy hình ảnh đưa vào bài bài báo do máy tính tự động xuất bản
https://beritagar.id/artikel/berita/sebagian-besar-sektor-menghijau-ihsgditutup-naik-110065
Danh mục những bài báo do máy tính tự động xuất bản https://beritagar.id/penulis/robotorial

Link bài viết: https://nguoilambao.vn/bao-chi-du-lieu-va-cong-nghe-tu-dong-lam-bao