**Biểu đồ thống kế người dùng và ngôn ngữ trong mạng xã hội tiềm ẩn**

Katja Filippowa

Google Inc.

Brandschenkestr. 110

Zurich, 8004 Switzerland

[katjaf@google.com](mailto:katjaf@google.com)

**Tóm tắc nội dung:**

Chúng tôi xem xét công việc dự đoán giới tính của người dùng YouTube và độ tương phản của 2 nguồn thông tin: Nội dung bình luận của họ và môi trường xã hội thể hiện từ đồ thị liên kết của người dùng và video. Chúng tôi tuyên truyền thông tin giới tính qua video và thấy rằng giới tính của một người dùng có thể dự đoán có độ chính xác trên 90%. Chúng tôi cũng thấy rằng giới tính có thể dự đoán được chỉ từ ngôn ngữ (89%). Một kết quả ngạc nhiên từ nghiên cứu này của chúng tôi là sau này dự đoán tương quan mạnh mẽ hơn với giới tính nổi bật trong môi trường người dùng hơn là giới tính của một người theo báo cáo trong hồ sơ cá nhân. Chúng tôi cũng khảo sát làm thế nào 2 khía cạnh (ngôn ngữ học và xã hội học) có thể được kết hợp và phân tích cách dự đoán chính xác thay đổi qua các nhóm tuổi khác nhau.

**1. Giới thiệu**

Trong vài thập kỷ qua, các trang web đã trở nên nhiều hơn và có nhiều mạng xã hội. Số người có một danh tính trong một trong các mạng xã hội (facebook, google+, twitter…) đã được tăng trưởng ổn định, nhiều người dùng giao tiếp trực tuyến với nhau mỗi ngày. Tương tác của họ mở ra những khả năng mới cho khoa học xã hội và ngôn ngữ học cũng không ngoại lệ. Ví dụ, với sự phát triển và lớn mạnh của web 2.0, nó sẽ có thể được truy cập vào khối lượng dữ liệu văn bản được gán nhãn đối với các thông số xã hội khác nhau như đất nước, tuổi, giới tính, chuyên môn hoặc tôn giáo. Nghiên cứu về bất đồng ngôn ngữ giữa nhóm ngăn cách bởi biến cố xã hội nhất định thuộc về lĩnh vực xã hội học trong đó nói chung điều tra tác động của xã hội về cách ngôn ngữ đã được sử dụng (Coulmas, 1998). Trong lịch sử, xã hội học được kết nối với ngôn ngữ địa phương tập trng chủ yếu vào khía cạnh ngữ âm của ngôn ngữ khu vực nhưng sau đó được mở rộng xã hội học (Chambers & Trudgill, 1998). Một nghiên cứu thông thường sẽ bao gồm lấy mẫu âm thanh từ dân số, phỏng vấn họ và phân tích các mục ngôn ngữ với sự liên quan tới thay đổi xã hội (Hudson, 1980).

Trong thập kỷ qua đã có nhiều nghiên cứu điều tra mối quan hệ giữa ngôn ngữ và thống kế dân số của người sử dụng blog hoặc twitter (xem Sec. 2 tài liệu tham khảo). Hầu hết các nghiên cứu sử dụng mạng xã hội để thu thập nhãn dữ liệu mẫu của văn bản cùng với thay đổi Demographics. Tuy nhiên, họ đã không phân thích làm các nào môi trường xã hội ảnh hưởng ngôn ngữ, mặc dù câu hỏi tương tự gần đây đã được đưa ra (nhưng chưa có câu trả lời) bởi Ellist (2009). Trong công việc của chúng tôi cố gắng để giải quyết chính xác vấn đề này. Đặc biệt, chung tôi coi công việc của dự đoán giới tính người dùng trên YouTube và tương phản hai nguồn thông tin: (1) bình luận được viết bởi người dùng và (2) khu vực xã hội lân cận như định nghĩa bởi đồ thị song song sử dụng video. Chúng tôi sử dụng bình luận để đào tạo phân loại giới tính trên một loạt các đặc điểm ngôn ngữ. Chúng tôi cũng giới thiệu một quy trình tuyên truyền giới tính đơn giản để dự đoán giới tính một người từ đồ thị người dùng video.

Trong phần tiếp theo, chúng tôi sẽ chứng minh rằng mặc dù ngôn ngữ cung cấp cho chúng tôi các tín hiệu biểu hiện giới tính người dùng (như báo cáo trong hồ sơ sử dụng), trong thực tế biểu hiện nhiều hơn một xã hội giới tính xác định. Bỏ qua những tranh luận về mối quan hệ phức tạp giữa ngôn ngữ và giới tính (xem Ecket & McConnell-Ginet 2003 cho cuộc thảo luận kỹ lưỡng của đối tượng), chúng tôi chỉ cần chứng minh rằng một phân lớp đào tạo để dự đoán giới tính trong môi trường mạng xã hội người dùng, như xấp xỉ bởi đồ thijYouTube của người sử dụng và video, đạt độ chính xác cao hơn cả hai giới tính so với một đào tạo để dự đoán giới tính bẩm sinh của người sử dụng. Chúng tôi cũng tìm hiểu cách làm thế nào quan điểm ngôn ngữ và xã hội có thể kết hợp để nâng cao độ chính xác của dự đoán. Cuối cùng, chúng tôi tìm thấy 3 nhóm: tuổi thanh thiếu niên, những người trong độ tuổi 20 và những người qua độ tuổi 30 và cho thấy rằng xác định giới tính rõ ràng hơn trong ngôn ngữ của người trẻ tuổi nhưng cũng có một sự tương phản cao giữa giới tính bẩm sinh và giới tính chủ đạo trong môi trường xã hội.

Bài báo được tổ chức như sau: Chung tôi sẽ điểm qua những công việc liên qua về ngôn ngữ của phương tiện truyền thông xã hội và Demographics người dùng (Sec. 2) và xây dựng mục tiêu nghiên cứu của chúng tôi (Sec. 3). Sau đó, chúng tôi miêu tả dữ liệu (Sec. 4) giới thiệu các thử nghiệm thống kế dân số (Sec. 5) và thử nghiệm giới tính có giám sát học máy từ ngôn ngữ (Sec. 6).

**2 Công việc liên quan**

Nghiên cứu trước đây về ngôn ngữ và thống kế dân số trong đó nhìn vào dữ liệu trực tuyến có thể phân biệt với sự tôn trọng các mục tiêu của họ. (1) Các nghiên cứu đến từ các mục đích cộng đồng xã hội học tại các giả thuyết thực nghiệm cá nhân, chẳng hạn rằng nữ nói sử dụng nhiều danh từ hoặc nam giới có xu hướng sử dụng nhiều từ dài. (2) Một mục tiêu chuẩn của nghiên cứu NLP là xây dựng một hệ thống tự động giải quyết một các chính xác một công việc nhấn định mà trong trường hợp thống kế dân số là dự đoán độ tuổi người sử dụng, giới tính hoặc quê hương. Trong phần này, chúng tôi bắt đầu bằng cách xem xét những kiểu nghiên cứu trước đó, trong đó có phân tích dữ liệu và kiểm tra giả thuyết. Ở đây thích hợp cho sự lựa chọn các đặc điểm của chúng tôi. Sau đó chúng tôi tóm tắc tổng kết của nghiên cứu về thống kế dân số dự đoán tốt hơn để xác định vị trí và trình bày cách tiếp cận của chúng tôi.

**2.1 Phân tích ngôn ngữ và Demographics**

Nghiên cứu xã hội học trước hầu như là kiểm tra giả thuyết được xây dựng trước khi sử dụng internet, chẳng hạn phụ nữ sử dụng hàng rào thường xuyên hơn (Lakoff, 1973) hoặc đàn ông sử dụng nhiều phủ định (Mulac et at, 2000) hoặc nhìn vào từ cụ thể hoặc các lớp từ. Newman et al (2008) cung cấp một đánh giá toàn diện về công việc như vậy và một mô tả về việc không sử dụng web trong tài liệu này. Một số giả thiết đã được khẳng định bằng thực nghiệm một số không có thực nghiệm.

Ví dụ, Herring & Paolillo (2006) phân tích giới tính và chi tiết giới tính sử dụng ngôn ngữ giao tiếp trực tuyến trên một mẫu khoảng 130 blog. Nhìn vào một số đặc trưng phong cách đã được đưa ra là dự đoán giới tính (Argamon et al, 2003;… Koppel et al, 2004), chẳng hạn đại từ nhân xưng, hạn định và chức năng từ loại khác, họ thấy không ảnh hưởng đến giới tính. Không giống họ, Kapidzic & Herring (2011) phân tích thông tin trò chuyện gần đây và xác định được giới tính. Tương tự, Hufaker & Calvert (2005) nghiên cứu câu hỏi về danh tính của bloggers thiếu niên (ví dụ, tuổi, giới tính, tình dục) và tìm thấy các đặc trưng ngôn ngữ biểu hiện của giới tính (ví dụ, sử dụng biểu tượng cảm xúc của đàn ông). Burger & Henderson (2006) xem xét mối quan hệ giữa ngôn ngữ (ví dụ, độ dài văn bản, sử dụng in hoa và dấu chấm câu) và đặc trưng phi ngôn ngữ (ví dụ, sở thích, tâm trạng) và độ tuổi của bloggers và vị trí địa lý. Họ thấy rằng nhiều đặc trưng phản ánh với tuổi tác và thử nghiệm với mục tiêu dự đoán xem liệu tác giả blog có thể 10 tuổi không.

**2.2 Demographics dự đoán từ ngôn ngữ**

Các nghiên cứu, chúng tôi xem xét ở đây sử dụng học máy có giảm sát để có mô hình để dự đoán giới tính hay tuổi tác. Thuộc tính Demographic khác nhau, như vị trí địa lý, dân tộc, hoặc trình độ giáo dục, cũng được dự đoán tự động (Gillick, 2010; Rao & Yarowsky, 2011, ngoài những điều khác). Như vậy, phương pháp sinh ra đã được áp dụng để phát hiện ra mối liên quan giữa ngôn ngữ và Demographics của người dụng phương tiện truyền thông xã hội (Eisenstein et al, 2011, ngoài những thứ khác) nhưng đây là liên quan trực tiếp ít hơn cho công việc hiện tại. Các phương pháp giám sát, các nguồn đặc trưng chính là các văn bản người dùng viết và cũng như hồ sơ các nhân liệt kê như tên, sở thích, bạn bè… Hiện này cũng đã có những nghiên cứu mà không xét đến ngôn ngữ tất cả nhưng được xem là môi trường xã hội mà thôi. Ví dụ, MackKinnon và Warren (2006) nhằm mục đích là dự đoán tuổi và vị trí của người dùng LiveJournal6. Những gì họ tìm thấy là có sự liên quan đáng kể giữa tuổi và vị trí của người dùng và những người bạn của họ, mặc dù có trường hợp ngoại lệ thú vị.

Burget et al (2011) đào tạo một phân loại giới tính trên tweet với từ và ký tự dựa trên đặc trưng ngram đạt độ chính xác 75.5%. Thêm các đặc trưng tên đầy đủ riêng cho phép tăng lên 89.1%, hơn nữa các đặc trưng như mô tả viết về chính mình và màn hình tên giúp hơn nữa để đạt được 92%. Ngoài ra, phương pháp tự đào tạo khám phá dữ liệu không có nhãn được mô tả nhưng hiệu suất thấp hơn. Các loại khác của đặc trưng xã hội học và một phân loại khác đã được áp dụng để dự đoán trên tweet bởi Rao và Yarowsky (2010).

Nowson và Oberlander (2006) đạt 92% độ chính xác các công việc dự đoán giới tính chỉ sử dụng đặc trưng ngram. Dữ liệu của họ gồm 1400/450 bài đăng viết bởi 47 nữ và 24 năm, tương ứng. Tuy nhiên, đặc trưng ngram được chọn trước dựa vào việc chúng xảy ra với tần số tương đối đáng kể trong ngôn ngữ của một giới trong giới khác. Kể từ khi tập dữ liệu hoàn chỉnh đã được sử dụng để chọn trước các đặc trưng, kết quả là không thuyết phục.

Yan & Yan (2006) đào tạo một phân loại Naïve Bayes để dự đoán giới tính của các tác giả blog. Trong tổng số họ tìm thấy 75000 mục blog cá nhân của 3000 tác giả blog, tất cả chúng đã đăng giới tính của họ trên trang cá nhân. Họ đo độ chính xác và hồi quy w.r.t. lớp thiểu số (nam) và nhận được f-measure tốt nhất của 0.64 (độ chính xác và hồi quy là 65% và 71% tương ứng).

Rosenthal và McKeown (2011) dự đoán tuổi của một blogger, đặc trưng hầu hết họ sử dụng chiết xuất từ các bài viết trên blog, đặc trưng khác bao gồm những điều blog thích, số bạn bè, thời gian thường đăng bài viết… Tương tự như vậy để Schler et al (2006), họ chạy một thử nghiệm phân loại với 3 lớp tuổi bỏ qua lứa tuổi trung niên và sử dụng các cơ sở lớp đa số để so sánh. Trong một thử nghiệm khác, họ thử nghiệm với một bộ phân loại nhị phân cho tuổi phân biệt thế hệ trước và sau truyền thông xã hội và sử dụng những năm từ 1975-1988 như một ranh giới. Độ chính xác của dự đoán tăng dần như theo những năm thực hiện.

Điều thú vị, nó đã thể hiện biểu đồ thống kế có thể dự đoán hạn chế nhiều hơn so với blog các nhân hoặc tweet và từ đoạn văn bản thậm chí đoán văn bản ngắn hơn tweet (Otterbacher, 2010; Popescu và Grefenstette, 2010).

**3 Động lực cho các nghiên cứu hiện này**

Tương tự nghiên cứu NLP trước đó, mục tiêu xuất phát của chúng tôi là dự đoán giới tính người dùng từ chính báo cáo của họ. Điều là đầu tiên của nghiên cứu của chúng tôi là làm như vậy, chung tôi đối chiếu hai nguồn dữ liệu thông tin: môi trường xã hội người dùng và văn bản được viết. Thực ra, một chủ đề mà chưa được nghiên cứu nhiều trong các nghiên cứu về ngôn ngữ và biểu đồ thống kê người dùng là mối quan hệ giữa ngôn ngữ của người dùng và môi trường xã hội. Các nghiên cứu phân tích dữ liệu đã xác nhận giả thiết liên quan đến sự phụ thuộc giữa đặc điểm ngôn ngữ (ví dụ, chiều dài trung bình của câu) và thông số biểu đồ thống kê (ví dụ, giới tính). Nghiên cứu dự đoán biểu đồ thống kế chủ yếu dự vào ngôn ngữ và đặc trưng hồ sơ người dùng một cách độc lập. Một ngoại lệ ở đây là Garera & Yarowsky (2009) chỉ ra rằng, dự đoán giới tính trong cuộc nói chuyện, nó giúp đỡ để biết giới tính của người đối thoại. Tuy nhiên, chúng tôi nhằm mục đích điều tra tác động của môi trường xã hội theo nghĩa rộng hơn nhiều so với người đối thoại trực tiếp và trong ngữ cảnh rộng hơn nhiều so với một cuộc trò chuyện.

Ngôn ngữ là một hiện tượng xã hội và nó thực tế là thúc đẩy tất cả nghiên cứu xã hội học. Nhiều ngôn ngữ nếu không phải tất cả đặc trưng ngôn ngữ không cứng nhắc hoặc bẩm sinh nhưng có thể giải thích bằng cách nhìn vào những người tương tác với nhau. Bởi vì mỗi ngôn ngữ nói có thể được xem như một thành viên của nhiều cộng đồng chồng chéo nhau (ví dụ, khoa học máy tính, Pháp, nam giới, vận động viên), ngôn ngữ của cá nhân có thể phản ánh mối quan hệ của họ trong các cộng đồng khác nhau với những mức độ khác nhau. Tương tác lặp lại với ngôn ngữ nói khác ảnh hưởng đến người nói (Baxter et al, 2006; Bybee, 2010) và sử ảnh hưởng quan sát được trên tất cả mức độ của đại diện ngôn ngữ (Croft, 2000). Ví dụ, nó được chỉ ra rằng càng có nhiều người thích hợp trong mộ cộng đồng nhất định và chặc chẽ hơn các mối quan hệ mạng xã hội, nổi bật là những đặc điểm đại điện cộng đồng trong ngôn ngữ cá nhân (Milroy & Milroy, 1992, Labov, 1994). Trong nghiên cứu, chúng tôi chấp nhận một quan điểm và phân tích ảnh hưởng của nó đối với dự đoán giới tính. Với bản chất xã hộ, không ngôn ngữ nào phản ánh các chuẩn mực của một cộng đồng người dùng trong đó hoặc giá trị thực của một biến đồ thị thống kê?

Trong nghiên cứu, chúng tôi giải quyết vấn đề với một mô hình kỹ thuật riêng: chúng tôi giả định rằng quan sát trực tiếp hành vi phản ánh tương ứng đời sống thực của người dùng (chi tiết hơn ở Sec 4 và 5) và dựa trên giả định này đưa ra kết luận về môi trường xã hội của người dùng. Chúng tôi sử dụng đặc trưng ngôn ngữ và một phương pháp giám sát để phân tích mối quan hệ giữa ngôn ngữ và các biến để dự đoán. Theo hiểu biết của chúng tôi, chúng tôi là người đầu tiên đặt câu hỏi liều đó thật sự là giới tính bẩm sinh mà phân loại ngôn ngữ dự vào để học dự đoán. Nhiều câu hỏi cụ thể, chúng tôi đang gợi ý trả lời như sau:

1. Những nghiên cứu trước đó đã xem xét dữ liệu trực tuyến dựa trên đồ thị thống kế của chính báo cáo. Hồ sơ dữ liệu được biết đến là không chuẩn mặc dù nó khó để đánh giá hồ sơ giả (Burger et al, 2011). Liên quan đến việc dự đoán, làm thế nào chúng ta có thể sử dụng những gì chúng ta biết về môi trường xã hội người dùng để làm giảm ảnh hưởng của độ chính xác (noise) Làm thế nào để chúng ta có lợi ích từ mẫu ngôn ngữ từ người dùng mà chúng ta chưa biết giới tính gì cả?

2. Khi phân tích ngôn ngữ của một người dùng, có bao nhiều đặc điểm giới tính cụ thể đúng để giới tính thực của người dùng và mức độ có thể họ được giải thích bởi môi trường xã hội? Sử dụng mô hình kỹ thuật của chúng tôi và một phân lớp ngôn ngữ nền tảng của giới tính, như thế nào là hiệu quả bị ảnh hưởng bởi những gì chúng ta biết về mội trường xã hội trực tuyến của người sử dụng?

3. Liên quan đến dự đoán giới tính giữa các nhóm tuổi khác nhau, làm thế nào để thay đổi phân loại hiệu suất? Đánh giá từ các giao tiếp trực tuyến, độ tuổi thanh thiếu niên có dấu hiện xác định giới tính nhiều hơn những người có tuổi? Xét về độ chính xác phân loại, dễ ràng để dự đoán giới tính của thanh thiếu niên hơn là của một người trưởng thành?

Sử mới lạ cuối cùng của nghiên cứu, chúng tôi là người đầu tiên cho thấy làm thế nào để YouTube có thể sử dụng được như là một nguồn tài nguyên có giá trị cho nghiên cứu xã hội học. Trong phần sau, chúng tôi làm nổi bật những đặc điểm làm cho YouTube thú vị và độc nhất.

**4 Dữ liệu**

Hầu hết các mạng xã hội đều cố gắng để bảo về sự riêng tư của người dùng và theo mặc định là không xem được những thông tin cá nhân (ví dụ, bài đăng, bình luận, bình chọn…). Để có được dữ liệu cho thực nghiệm của chúng tôi, chúng tối sử dụng YouTube, một trang web chia sẻ video. Hầu hết người đăng kí dùng YouTube đều có giới tính, tuổi và vị trí trên trang hồ sơ cá nhân của họ, cũng như bình luận đều được công khai. YouTube là một trang thú vị cho các nghiên cứu xã hội học vì nhiều lý do:

- Tính đa dạng cao: Nó không giới hạn cho bất kỳ chủ đề nào (ví dụ, những blog về chính trị) nhưng bao gồm một hàng loạt các chủ đề thu hút nhiều người, từ trẻ em quan tâm đến phim hoạt hình cho đến các học giả xem bài giảng về triết học.

- Lời nói tự phát: bình luận của người dùng được cho là nhiều tự phát hơn blog có nhiều khả năng để phù hợp với chuẩn mực của ngôn ngữ viết. Đồng thời họ hạn chế hơn các tweet được viết theo hạn chế chiều dài trong đó khuyến khích các phát ngôn ngắn gọn xúc tích.

- Dữ liệu có sẵn: Tất cả bình luận đều công khai, vì vậy chúng tôi có không lấy được một tập hợp con ảnh hưởng của những gì một người dùng viết công khai. Hơn nữa, chúng tôi quan sát tương tác của người dùng trong những môi trường khác nhau bởi vì mọi video nhắm tới nhóm người rõ ràng những người có thể chia sẽ nguồn gốc (ví dụ, các cuộc bầu cử ở Hy Lạp) hoặc sở hữu (ví dụ, làm thế nào để mở khóa iPhone) hoặc bất cứ tài sản khác. Một số video thu hút một nhóm người xác định (ví dụ, gia đình của một đứa trẻ mới sinh) trong đó một số video thu hút một lượng khán giả rất rộng (ví dụ, video về một con meo con).

***Bảng 1: Phân bổ giới tính từ 6.9 triệu người dùng***

Từ quan hệ người dùng, video và bình luận chúng tôi xây dụng một biểu đồ liên kết (Easley & Kleinberg, 2010): một người dùng và một video liên kết nếu người dùng bình luận video đó (Hình 1(a)). Đồ thị của chúng tôi là một trọng số mặc dù số lượng bình luận có thể được sử dụng để tăng cạnh. Đồ thị đồng bình luận là một phiên bản chặt chẽ của đồ thì đồng quan điểm phổ biến hơn được sử dụng, ví dụ, các nghiên cứu khuyến cáo video (Baluja et al, 2008, ngoài những điều khác).

Chúng tôi thu được một mẫu ngẫu nheien của video bằng các xem xét tất cả các video mà YouTube có prefix8 cụ thể. Từ đó, chúng tôi thu thập hồ sơ của người dùng nhận xét trên video. Tổng cộng, chúng tôi lấy được 6.9 triệu hồ sơ người dùng có viết ít nhất 20 bình luận, không nhiều hơn 30 bình luận được thu thập với mỗi người dùng. Ngưỡng về số lượng tối thiểu trên bình luận được thiết lập để giảm tỷ lệ người dùng đã sử dụng YouTube chỉ một vài lần và có thể theo gợi ý của trang web trong sự lựa chọn video của họ. Phân bố giới tính người dùng trình bày trong bảng 1. Mặc dù nữ ở độ tuổi thanh thiếu niên được báo cáo là có nhiều khả năng để viết blog hơn là nam giới (Herring et al, 2004), nam giới chiếm ưu thế trong tập dữ liệu. Một mẫu ngẫu nhiên từ một người dùng chung bao gồm có ngưỡng 20 bình luận thể hiện rằng có nhiều nam giới hơn tổng thể, mặc dù sự khác biệt không đáng kể đối với thanh thiếu niên: 58% thanh thiếu niên biết giới tính nam ngược lại với 74% và 79% cho các nhóm 20-29 và 30 tuổi. Thanh thiếu niên cũng chiếm nhiều hơn khoảng 35% trong dữ liệu chúng tôi.

Mặc dù chúng tôi không lọc người dùng dựa trên vị trí hoặc tiếng mẹ đẻ như nhiều người bình luận trong nhiều ngôn ngữ, tập bình luận chủ yếu là tiếng anh.

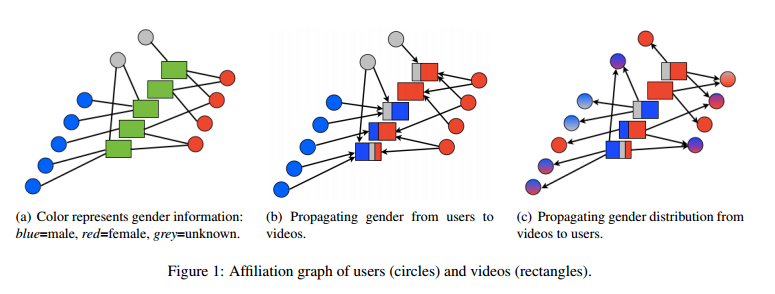
**5 Tuyên truyền giới tính**

Chúng tôi đầu tiên xem xét môi trường xã hội của người dùng để xem liệu có bất kỳ mối liên hệ giữa giới tính của mội người dùng và phân bố giới tính trong khu vực xung quanh, độc lập với ngôn ngữ. Chúng tôi sử dụng một thủ tục tuyên truyền đơn giản để đạt được những người hàng xóm gần nhất của người dùng, có nghĩ rằng, những người khác “thuộc” với cùng video. Cụ thể, chúng tôi thực hiện theo 2 bước:

1. Chúng tôi gửi thông tin giới tính (nữ, nam, chưa biết) tới tất cả video người dùng có bình luận. Bằng cách này mỗi video chúng tôi có được một phân bổ đa dạng trên 3 lớp (xem hình 1(b)).

2. Chúng tôi gửi bản phân bố giới tính từ mọi video tới tất cả người dùng bình luận trên nó và trung bình trên tất cả video người dùng kết nối (xem hình 1(c)). Tuy nhiên, trong khi làm như vậy chúng tôi điều chỉnh phân bổ cho mỗi người dùng để biểu đồ thông kế được loại trừ. Bằng cách này chúng tôi có một thiết lập công bằng nơi giới tính ban đầu cảu người dùng không bao giờ có trong những gì lấy lại từ kết nối video. Như vậy, giới tính của người dùng góp phần vào sự phân bổ kề cận với tất cả vùng xung quanh nhưng không phân bổ giới tính cuối cùng của mình.

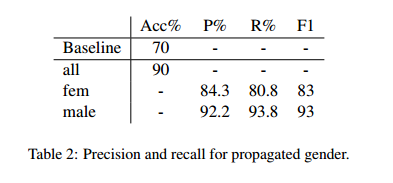
Cùng với động lực của chúng tôi và kỹ thuật mô hình, chúng tôi đã chọn một phương pháp đơn giản (và không, nói, phân loại) để xấp xỉ các cuộc gặp gỡ trực tiếp của người dùng: Thường gặp gỡ phụ nữ hoặc đàn ông? Theo cách này chúng tôi nghĩ rằng các video tương ứng tới một thăm những nơi( ví dụ, một rạp chiếu phim, một hãng mỹ phẩm, một quán rượu) bởi người dùng nơi vô ý hay cố ý tiếp xúc với các người nói sử dụng ngôn ngữ. Tương tự như Baxter et al (2006), chúng tôi giả định rằng những cuộc gặp gỡ ảnh hưởng đến cách người đó nói. Lưu ý rằng nếu giới tính người dùng không ảnh hưởng đến sự lựa chọn video, sau đó, trên trung bình, chúng tôi mong đợi mỗi video để có sự phân bổ tương tự như trong dữ liệu của chúng tôi nói chung: 62% nam, 26% nữ, 12% không rõ (Bảng 1).



Để có được một dự đoán giới tính duy nhất từ phân bổ tuyên truyền, đối với một người dùng nào đó, chúng tôi chọn lớp giới tính (nữ hoặc nam) trong đó có nhiều hơn các phân bổ tập trung. Các thủ tục chính xác là như sau: Cho người dùng u kết nối với video , có m giới tính phân bổ gửi tới u: . Một phân bổ duy nhất lấy từ

Để giải quyết các độ lệch trong dữ liệu, ví dụ, thực tế hơn 70% người dùng của chúng tôi (62/(26+62)) biết giới tính là nam, chúng tôi lựa chọn giới tính nữ nếu (a) có nhiều hơn 0 và ít nhất là nhiều hơn nam: , hoặc (b) nó có ít nhất *τ* của tập thể: . Chúng tôi thiết lập *τ = 0.26* ban đầu bởi vị nó tương ứng với tỷ lệ dự kiến của nữ (26%) nhưng tiếp tục thử nghiệm với các giá trị *τ* trong khoảng 0.25 – 0.4. Chúng tôi lấy được độ chính xác tốt nhất và f-measures với ngưỡng 0.33, sự khác nhau về độ chính xác từ ngưỡng ban đầu 0.26 là ít hơn 2%. Thực tế rằng, giá trị *τ* tối ưu là khác với tỷ lệ chung của nữ (26%) là không phải ngạc nhiên vì chúng tôi thống kế mỗi phân bổ video và không đếm dữ liệu chưa xử lý của người dùng.

Những dự đoán thu được với các phương pháp tuyên truyền được mô tả khá chính xác đạt độ chính xác là 90% (Bảng 2). Cơ sở giao tất cả những lớp người dùng (tất cả nam) cung cấp cho chúng tôi với độ chính xác 70% - tỷ lệ nam trong số những người sử dụng biết giới tính).



Mặc dù mục đích của phần này không phải để trình bày một phương pháp dự đoán giới tính, chúng tôi thấy nó có giá trị nhấn mạnh rằng độ chính xác 90% là rất đang quan tâm vì chúng ta chỉ nhìn vào người dùng lân cận. Trong phần sau, chúng tôi sẽ điều tra cách nhìn xã hội trên biểu đồ thống kê có thể giúp chúng tôi dự đoán giới tính từ ngôn ngữ.

6 Học máy có giám sát giới tính

Trong phần này chúng tôi bắt đầu bằng mô tả thử nghiệm dự đoán giới tính dầu tiên và một vài mở rộng và sau đó lần lượt đến các kết quả.

**6.1 Thử nghiệm**

Tương tự như nghiên cứu trước đó dự trên dự đoán biểu đồ thống kế, chúng tôi bắt đầu với phương pháp giám sát và chỉ quan sát đoạn văn bản (bình luận) được viết bởi người dùng. Chúng tôi không dựa trên bất kỳ thông tin gì từ môi trường xã hội của người dùng và không sử dụng bất kỳ đặc trưng mở rộng từ hồ sơ người dùng, giống như tên, cái mà có thể được dùng cho việc dự đoán giới tính dễ dàng hơn (Burger et al, 20111). Cuối cùng, chúng tôi không mở rộng bất kỳ đặc trưng nào từ video người dùng bình luận bởi vì mục tiêu của chúng tôi ở đây là để khám phá ngôn ngữ như một nguồn thông tin duy nhất. Ở đây, chúng tôi chỉ đơn giảm muốn điều tra mức mộ mà các ngôn ngữ của người dùng là biểu hiện của giới tính mà được tìm thấy trong hồ sơ và trong đó bỏ qua thông tin không rõ ràng, tương ưng với giới tính bản sinh.

Trong thử nghiệm, chúng tôi sử dụng phân bổ thực hiện của học tối đa entropy (Berger et al 1996; McDonald et al, 2010) mà đầu ra một phân bổ qua các lớp, dự đoán cuối cùng là lớp với xác suất lớn nhất. Chúng tôi lấy 80% của người dùng để đào tạo và tạo ra một thực thể đào tạo cho mỗi người dùng thực hiện giới tính thể hiện trên trang cá nhân (4.9 triệu) Còn lại 20% của dữ liệu là sử dụng để kiểm tra (1.2 triệu). Chúng tôi sử dụng theo 3 nhóm đặc trưng: (1) dựa trên ký tự: độ dài trung bình của bình luận, tỉ lệ chữ hoa trong tổng số chữ, tỷ lệ dấu chám trong tổng số kí tự; (2) dấu hiện: trung bình độ dài bình luận trong từ, tỷ lệ của các từ duy nhất trong tổng dẫu hiệu, chữ thường unigram với tổng số trên các ý kiến (10k thường xuyên unigram sử dụng, tần số tính toán trên tập bình luận riêng biệt) sử dụng các đại từ, từ hạn định, từ chức năng; (3) trên câu: độ dài trung bình của bình luận trong câu, độ dài trung bình của câu trong từ.

**Tăng cường tập huấn luyện**: Câu hỏi đầu tiên chúng tôi xem xét là làm thế nào đồ thị liên kết và tuyên truyền giới tính có thể được sử dụng để nâng cao dữ liệu của chúng tôi cho các thử nghiệm giám sát. Một khả năng có thể để huấn luyện một phân loại trên một tập đã điều chỉnh của người sử dụng bằng cách loại bỏ tất cả những người mà báo cáo giới tính không phù hợp với giới tính dự đoán của vùng lân cận. Điều này có lẽ sẽ làm giảm lượng tiếng ồn bằng cách loại bỏ những người dùng cố tình cung cấp thông tin sai sự thật về hồ sơ của họ. Một khả năng sẽ được mở rộng việc huấn luyện với những người dùng không hiển thị giới tính công khai nhưng chúng tôi có thể dự đoán gần đúng. Ý tưởng ở đây là tương tự như cùng huấn luyện, nơi đã có một trong hai quan điểm trên dữ liệu. Trong phần này, một đồ thị xã hội sẽ được kết hợp với quand diểm dựa trên ngôn ngữ.

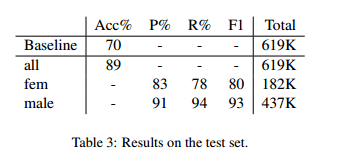
**Hồ sơ cá nhân với dự đoán giới tính lân cận**. Câu hỏi tiếp theo đặt ra trong phần động lực như sau: Có một thực tế rằng ngôn ngữ là một hiện tượng xã hội và nó đã được định hình bởi môi trường xã hội của người nói ảnh hưởng tới phân loại giới tính? Nếu thật sự có đặc điểm ngôn ngữ chỉ rõ giới tính và được phản ánh trong các đặc trưng của chúng tôi thì chúng ta không thấy bất kỳ sự khác biệt có ý nghĩ gì giữa những kết quả dự đoán người dùng phù hợp với giới tính di truyền từ lân cận và những người có giới tính không phù hợp. Một giả thiết ngược lại sẽ là những điều phân lớp thật sự học để dự đoán không giống như bẩm sinh mà là một giới tính xã hội. Trong trường hợp này, phân lớp huấn luyện trên nhãn giới tính di truyền nên sẽ chính xác hơn một huấn luyện trên nhãn lấy từ hồ sơ cá nhân.

Để giải quyết những câu hỏi, chúng ta so sánh 2 phân loại: (1) Một mô tả trong đoạn đầu của phần này được huấn luyện trên tập nhãn giới tính từ hồ sơ cá nhân người dùng; (2) một phân loại huấn luyện trên giới tính lân cận, đó lá giới tính chi phối của môi trường của người nói có được với các thủ tục mô tả trong phần 5.

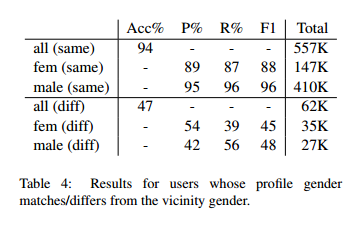
Nhóm độ tuổi và dự đoán giới tính. Cuối cùng, chúng ta nhìn các dự đoán giới tính thay đổi với độ tuổi và huấn luyện 3 mô hình độ tuổi chi tiết đẻ dự đoán giới tính với thanh thiếu niên (13-19), những người ở độ tuổi đôi mươi (20-29) và những người trên 30 (30+), độ tuổi cũng được lấy từ hồ sơ cá nhân. Những nhóm này được xác định để kiểm tra liệu thanh thiếu niên có xu hướng làm nổi bật đặc điểm giới tính hơn những người nhiều tuổi hơn không, một giả thiết nghiên cứu trước đây trên dữ liệu mẫu những bài viết của blog (Huffaker & Calvert, 2005).

6.2 Kết quả

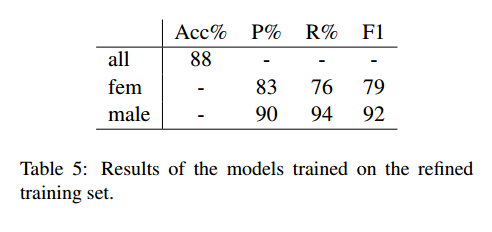
Chúng tôi báo cáo kết quả của thí nghiệm giám sát cho tất cả các thiết lập được mô tả ở trên. Như một đánh giá các ràng buộc thấp nhất chúng tôi cũng đưa ra kết quả của lớp đa số cơ bản (tất cả nam) đảm bảo độ chính xác 70%. Đối với phân loại giám sát chúng tôi báo cáo chính xác và độ chính xác giới tính, recall và fmeasure. Table 3 thể hiện kết quả bắt đầu phân loại huấn luyện để dự đoán giới tính hồ sơ cá nhân.



Để điều tra mối quan hệ giữa môi trường xã hội của một người, giới tính và ngôn ngữ, chúng tôi phân tách người dùng từ tập kiểm tra vào hai nhóm: Những người có giới tính phù hợp với giới tính di truyền từ vùng lân cận và những người cho kết quả không phù hợp. Như vậy Table 4 thể hiện kết quả giống như Table 3 nhưng tách cho hai nhóm người dùng. Nó cũng cho người dùng đếm w.r.t giới tính hồ sơ cá nhân.



Tăng cường tập huấn luyện. Trong thí nghiệm tiếp theo, chúng ta cải thiện tập huấn luyện bằng các loại bỏ tất cả người dùng có giới tính lân cận không phù hợp với giới tính báo cáo trong hồ sơ cá nhân. Việc đánh giá được làm trên tập không thay đổi (Table 5). Dự đoán được lấy từ mô hình huấn luyện trên một tập đã rút trích của người dùng đưa ra được một ít chính xác hơn so với những người thực hiện bởi các mô hình đào tạo trên tập huấn luyện đầy đủ (Table 3). Các mô hình rút trích thực hiện nhẹ (<1%) tốt hơn so với bắt đầu một người sử dụng vùng lân cận và giới tính hồ sơ phù hợp nhưng có kết quả rất kém trên người sử dụng với một giới tính không phù hợp, độ chính xác thấp như 37%. Độ chính xác bắt đầu mô hình trên người dùng là 47% (Table 4).



Trong thí nghiệm khác chúng tôi mở rộng dữ liệu huấn luyện với người dùng không biết giới tính nhưng dự đoán được với phương pháp tuyên truyền. Tuy nhiên, một tập dữ liệu lớn làm ra một khác biệt chỉ khi có một hiểu suất cao hơn so với kích thước ngày càng tăng của tập huấn luyện. Chúng ta quan sát chỉ thấy tăng nhẹ trong thực hiện (<1%) khi kích thước dữ liệu huấn luyện đã tăng một bậc. Do đó, không ngạc nhiên rằng việc thêm 12% không ảnh hưởng đến kết quả.

Ngôn ngữ, lân cận và giới tính hồ sơ cá nhân. Khoảng cách về độ chính xác của dự đoán với 2 nhóm người dùng trong Table 4 là đáng chú ý: 47% và 94%.