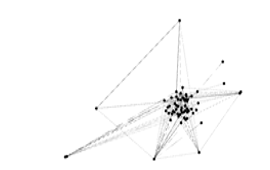
**I, MỞ ĐẦU**

**Giới thiệu**

Với sự phát triển mạnh mẽ của internet toàn cầu như hiện nay, nhu cầu trao đổi, tìm kiếm thông tin… online ngày càng bùng nổ. Đặc biệt, sự ra đời của các trang mạng xã hội với số lượng người dùng đông đảo trên toàn thế giới đã trở thành một nơi giao lưu kết bạn, trò chuyện và chia sẻ thông tin không thể thiếu của bất kỳ cá nhân, tổ chức nào. Do đó, mạng xã hội là một nơi lưu trữ thông tin và các mối quan hệ khổng lồ. Việc phân tích cấu trúc mạng xã hội để tìm hiểu và khai thác hiệu quả kho thông tin hữu ích này sẽ đem lại cho chúng ta rất nhiều lợi ích trong các lĩnh vực: *xã hội, kinh tế, y tế, quốc phòng-an ninh…* Do đó, *phân tích cấu trúc mạng xã hội ứng dụng vào bài toán thực tế phân tích tâm lý người dùng* đã và đang trở thành một đề tài mới mẻ, hấp dẫn trong những năm gần đây.

Các trang mạng xã hội phổ biến hiện nay như: *Facebook, twitter, Google+, …* Mạng xã hội là dịch vụ kết nối các thành viên cùng sở thích trên Internet lại với nhau với nhiều mục đích khác nhau mà không phân biệt không gian và thời gian. Bất kỳ một website nào mang tính chất cộng đồng, xây dựng nhằm mục tiêu thu hút người dùng *internet* tham gia dựa trên một đặc điểm chung về sở thích nào đó thì đều được coi là mạng xã hội, ví dụ như: Forum, website chia sẻ hình ảnh, video hay blog… cũng đều là các mạng xã hội. Mạng xã hội thường được định nghĩa như một đồ thị khổng lồ với các đỉnh là các thành viên và các cạnh là các liên kết thể hiện mối quan hệ hay sự tương tác giữa các thành viên đó.



Hình 1: Mô hình đồ thị mạng xã hội.

Luận văn sử dụng mạng xã hội của các thành viên Wikipedia để tiến hành thu thập và phân tích các bình luận, đánh giá mà các thành viên giành cho nhau trong các cuộc tranh cử vị trí quản trị viên. Wikipedia nói riêng và Wiki nói chung là một chương trình máy chủ cho phép người dùng hợp tác trong việc hình thành nội dung của một trang web. Wikipedia chỉ là một trong số rất nhiều trang web phát triển từ nền tảng wiki như: Wikileaks, Wikigames… có cộng đồng phát triển đông đảo và không ngừng lớn mạnh theo thời gian.

Đề tài nghiên cứu này tập trung chủ yếu vào bài toán *dự đoán loại liên kết*, cụ thể là bài toán dự đoán loại liên kết “positive” và “negative” trong mạng xã hội có dấu. Trong thực tế, quan điểm đánh giá giữa người này với người khác được sử dụng rộng rãi trong các bài diễn văn, bài luận, cuộc đàm thoại và luôn hàm chứa hai loại liên kết: positive (mối quan hệ tích cực giữa bạn bè hoặc những người cùng quan điểm) và negative (mối quan hệ tiêu cực giữa hai đối tượng có quan điểm đối lập). Do đó, việc dự đoán quan điểm người dùng là yếu tố quan trọng được ứng dụng trong: *việc xây dựng danh tiếng khi tranh cử, xây dựng các liên kết xã hội và khảo sát lấy ý kiến công chúng*.

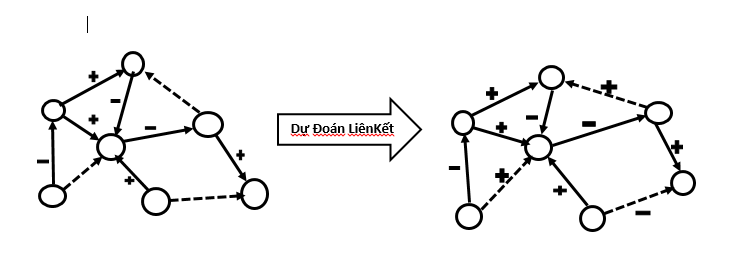
Trong các phương pháp cổ điển, việc đánh giá quan điểm người dùng có thể được phân tích riêng biệt bằng cách sử dụng mạng xã hội có dấu hoặc phân tích tâm lý dựa trên nội dung văn bản (sử dụng máy học là một phần của xử lý ngôn ngữ tự nhiên). Tuy nhiên, các phương pháp này đã bỏ qua sự tương tác phong phú giữa ngôn ngữ và bối cảnh xã hội. Đề tài được thực hiện dựa trên phương pháp mới được đề xuất bởi *Robert* có thể nắm bắt được các tương tác như vậy bằng cách tổng hợp các thông tin từ mạng xã hội có dấu như: *mối quan hệ giữa các người dùng trong mạng xã hội và các văn bản đánh giá liên quan*.

**Phát biểu bài toán**

Mạng xã hội là một mô hình mạng (đồ thị) G được cấu tạo bởi các đỉnh và các cạnh:

* Các đỉnh là tập đối tượng V.
* Các cạnh là tập các liên kết E.

Luận văn quy định rằng liên kết “*positive*” sẽ được biểu diễn bằng một cạnh có dấu “+” trên đồ thị G và ngược lại liên kết “*negative*” sẽ được biểu diễn bằng một cạnh có dấu “−”.



Hình 2: Dự đoán liên kết trong mạng xã hội.

Bài toán phân tích cấu trúc mạng xã hội trên *Wiki* kết hợp giữa phương pháp phân tích mạng xã hội có dấu (gồm hai trạng thái là “*positive*” và “*negative*”) và phân tích tâm lý dựa trên nội dung văn bản có *input* và *output* như sau:

* ***Input***: Thông tin dấu hiệu cạnh có sẵn của đồ thị mạng xã hội Wikipedia tại một thời điểm cụ thể và đoạn văn bản text liên quan là lời bình luận, nhận xét của các thành viên.
* ***Output***: Quan điểm của người dùng A với người dùng B là tích cực hay tiêu cực.

**Ứng dụng của bài toán**

Mạng xã hội của *Wiki* có đầy đủ các đặc trưng của một mạng xã hội thông thường: là một đối tượng động, có cấu trúc thay đổi thường xuyên theo thời gian. Do đó, bài toán này khi được giải quyết thành công sẽ dự đoán được cấu trúc mạng xã hội trong tương lai để ứng dụng vào thực tế như:

* Xác định cấu trúc mạng lưới cử tri và dự đoán quan điểm của họ. Từ đó, có chiến dịch tranh cử thích hợp nhằm tăng cường danh tiếng trong các cuộc bỏ phiếu tín nhiệm, tranh cử, bình chọn.
* Ứng dụng trong lĩnh vực chính trị nhằm khảo sát ý kiến người dân về các chính sách của Nhà nước.
* Cải thiện chức năng giới thiệu bạn mới phù hợp trên mạng xã hội (kết nối những cá nhân có chung quan điểm về một vấn đề nào đó).
* Các hệ thống tìm kiếm nâng cao.

Nhận thấy, bài toán *phân tích cấu trúc mạng xã hội vào phân tích tâm lý người dùng* là một trong những bài toán thu hút sự quan tâm rất lớn hiện nay vì khả năng ứng dụng thực tế cao và tính mới mẻ của nó.

**Khó khăn và thách thức của bài toán**

Phương pháp tiếp cận của bài toán dự đoán quan điểm người dùng trên mạng xã hội mà luận văn sử dụng là một phương pháp tiếp cận khá mới mẻ so với các phương pháp truyền thống (chỉ sử dụng mạng xã hội có dấu hoặc phân tích cảm xúc bằng văn bản). Việc kết hợp hai phương pháp đơn lẻ vào cùng một hệ thống sẽ tăng độ chính xác và độ tin cậy của kết quả thu được, cũng như thể hiện được mối quan hệ giữa bối cảnh xã hội và quan điểm người dùng. Tuy nhiên, bên cạnh đó, đề tài cũng gặp những khó khăn và thách thức nhất định khi xem xét mạng xã hội Wiki trong bối cảnh có hai loại liên kết là “*positive*” và “*negative*”. Chẳng hạn như:

* *Mạng xã hội có tính biến động và độ lớn cao*: Mạng xã hội là loại dữ liệu có tính biến động cao và có khả năng bùng nổ về kích thước cao với số lượng thành viên lớn. Ngoài ra, số lượng liên kết được thêm vào và bỏ đi cũng luôn luôn có sự thay đổi không ngừng làm thay đổi cấu trúc mạng gây khó khăn cho việc nghiên cứu.
* *Sự chênh lệch về quan điểm “positive” và “negative” trong mạng xã hội*: Thông thường, mạng xã hội luôn có tỷ lệ “*positive*” chiếm đa số, làm mất cân bằng mạng, dẫn đến sai lệch kết quả phân tích nghiêng về “*positive*”. Do đó, luôn cần một chi phí nhất định để xử lý sự chênh lệch này.

**Các công trình nghiên cứu liên quan**

[1] R. West, H. S. Paskov, J. Leskovec, C. Potts, “Exploiting Social Network Structure for Person-to-Person Sentiment Analysis”, in *Transactions of the Association for Computational Linguistics (TACL), 2,* 2014, eprint arXiv:1409.2450.

[2] Anh-Thu Nguyen-Thi, Phuc Quang Nguyen, Thanh Duc Ngo, Tu-Anh Nguyen-Hoang, “Transfer AdaBoost SVM for Link Prediction in Newly Signed Social Networks using Explicit and PNR Features”, in *Proceedings of The Nineteenth International Conference on Knowledge-Based and Intelligent Information & Engineering Systems*, 7-9/9/2015, Singapore.

[3] J. Leskovec, D. Huttenlocher, J. Kleinberg, “Predicting positive and negative links in online social networks”, n *WWW*, 2010, p. 641-650.

[4] J. Leskovec, D. Huttenlocher, J. Kleinberg, “Signed networks in social media”, in *Proc. 28th CHI*, 2010.

**Framework**

Đặc trưng explicit (vector 9 chiều): Thể hiện tính chất đặc trưng của liên kết e = (u, v):

* *Độ đo của đỉnh*: Thể hiện mối quan hệ cục bộ của một đỉnh với các đỉnh còn lại trong đồ thị.
* *Độ đo trung tâm của đỉnh*: Thể hiện tầm quan trọng của đỉnh đó trong đồ thị.
* *Số lượng bộ ba*: Được xây dựng dựa trên lý thuyết cân bằng thuộc về tâm lý xã hội.

Đặc trưng PNR (vector 2 chiều): Vận dụng các quy luật hành vi và tâm lý xã hội.

* *Tâm lý bầy đàn*: Một người có nhiều liên kết đến là liên kết *positive* thì có xu hướng nhận thêm các liên kết *positive* và ngược lại.
* *Thói quen và hành vi ra quyết định:*Một người có nhiều liên kết ra là liên kết *positive* thì có xu hướng tạo thêm nhiều liên kết *positive* và ngược lại.
* *Quy luật nhân quả*: Một người đã tạo nhiều liên kết ra là liên kết *positive* thì có xu hướng sẽ nhận liên kết *positive* và ngược lại.

**Dataset**

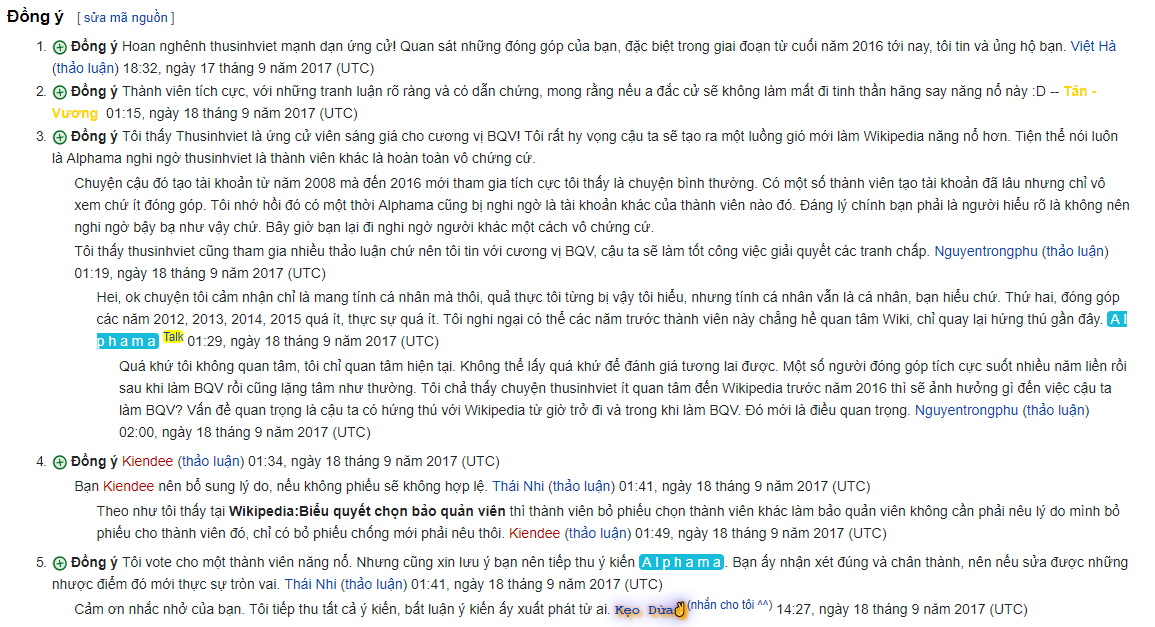
Dataset là bộ dữ liệu được thu thập từ mạng xã hội giành cho thành viên của Wikipedia do Đại học StanFord (Hoa Kỳ) công bố với mục đích đánh giá khả năng dự đoán của mô hình đánh giá quan điểm của các thành viên trên các bài viết không chính thức trên web nhưng liên quan đến các kết quả xã hội quan trọng.

Một thành viên của cộng đồng Wikipedia muốn trở thành quản trị viên thì một yêu cầu cho adminship (RFA) là phải nộp đơn hoặc tham gia ứng cử để những thành viên khác của cộng đồng wikipedia bầu chọn theo 3 quan điểm: ủng hộ (supporting), đối lập (neutral), trung lập (oppose). Bộ dữ liệu thu thập và phân tích trong các quá trình RFA từ năm 2003 đến tháng 05/2013 trên tất cả các phiếu bầu. Bộ dữ liệu gồm 11.042 người dùng được tạo thành 189.004 cặp cử tri riêng biệt với 198.275 phiếu bầu, mỗi phiếu bầu sẽ kèm theo một đoạn text bày tỏ quan điểm (một ứng viên có thể tham gia vận động tranh cử cho vị trí Adminship nhiều lần và mỗi cặp cửu tri có thể đóng góp ý kiến nhiều lần).

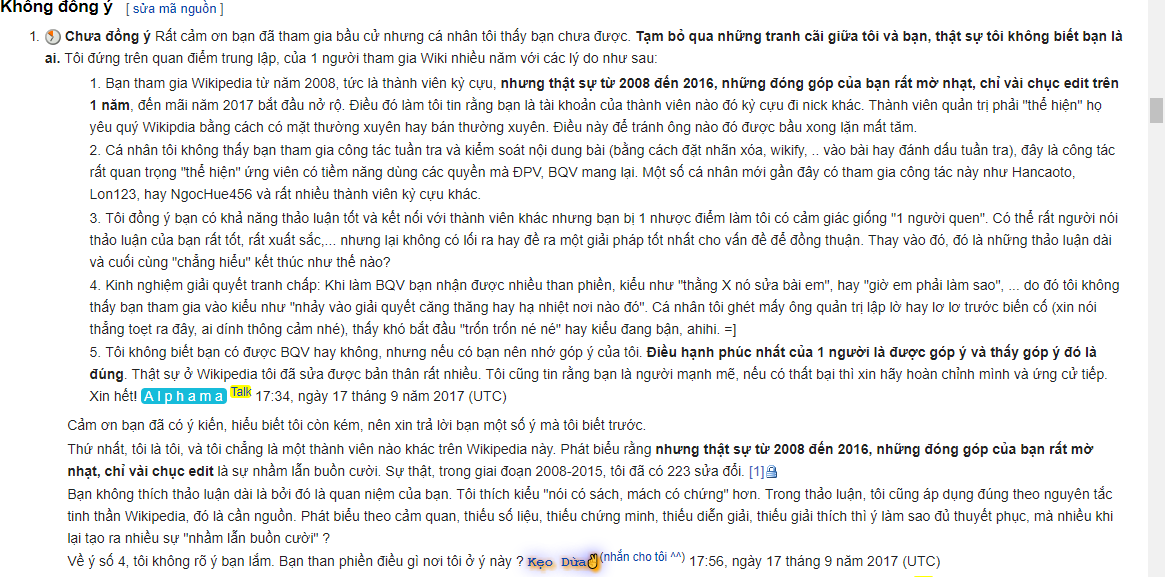
Dataset được thu thập trên 6 trường, đó là:

* SRC: tên cử tri
* TGT: tên người tham gia tranh cử
* VOT: nguồn vote chủ yếu(-1 = phản đối; 0 = trung tính; 1 = hỗ trợ)
* RES: kết quả cuộc bầu cử (-1 = bị từ chối; 1 = được chấp nhận)
* YEA: năm cuộc bầu cử diễn ra
* TXT: nhận xét bằng văn bản của các nguồn, trong wiki

Chúng ta có thể dễ dàng gia nhập vào mạng xã hội này của Wikipedia sau khi đăng ký thành viên và đăng nhập bằng tên và mật khẩu đã đăng ký. Dưới đây là ví dụ về vấn đề tranh cử Adminship trong Wikipedia của một thành viên ThưsinhViet.



Hình 3: Mục bình luận theo quan điểm ủng hộ



Hình 4: Mục bình luận theo quan điểm đối lập.