Phát hiện tài khoản spam trên mạng xã hội dựa trên phương pháp lai (Một thực nghiệm)

Vương Thị Hồng
Đại học Công nghệ, Đại học Quốc gia Hà Nội
UET-VNU
Hà Nội, Việt Nam
hongvt_57@vnu.edu.vn

Trần Văn Hiến
Đại học Công nghệ, Đại học Quốc gia Hà Nội
UET-VNU
Hà Nội, Việt Nam
hientv_55@vnu.edu.vn

Nguyễn Minh Đức
Đại học Công nghệ, Đại học Quốc gia Hà Nội
UET-VNU
Hà Nội, Việt Nam
ducnm_57@vnu.edu.vn

Nguyễn Thị Cẩm Vân
Đại học Công nghệ, Đại học Quốc gia Hà Nội
UET-VNU
Hà Nội, Việt Nam
vanntc_58@vnu.edu.vn

Nguyễn Văn Quang
Đại học Công nghệ, Đại học Quốc gia Hà Nội
UET-VNU
Hà Nội, Việt Nam
quangnv_570@vnu.edu.vn

Trần Mai Vũ
Đại học Công nghệ, Đại học Quốc gia Hà Nội
UET-VNU
Hà Nội, Việt Nam
vutm@vnu.edu.vn

***Tóm tắt*—Mạng xã hội trên nền tảng web ngày càng thu hút một số lượng lớn người dùng tham gia kết nối, tương tác và chia sẻ với nhau. Sự phát triển mạnh mẽ của mạng xã hội như Facebook kéo theo sự tăng lên nhanh chóng của tài khoản spam cả về quy mô lẫn tính chất. Những tài khoản spam này thường tiến hành bằng các cuộc tấn công lừa đảo, phần mềm độc hại và tin nhắn spam thương mại. Tài khoản spam đăng bài hoặc bình luận trên các trang (Page) để gửi nội dung thông điệp spam tới bạn bè của họ hoặc bạn bè của người khác trên mạng Facebook. Trong bài báo này, chúng tôi giải quyết bài toán nhận dạng tài khoản spam trên mạng xã hội Facebook dựa trên nội dung bình luận và hành vi người dùng. Chúng tôi đề xuất một phương pháp lai sử dụng mô hình Maximum Entropy cho bài toán phân lớp bình luận có phải là spam hay không. Chúng tôi đã tiến hành thử nghiệm tập dữ liệu thu thập được trên mạng xã hội Facebook để xây dựng mô hình phát hiện tài khoản spam và thu được những kết quả khả quan ban đầu. Kết quả trung bình độ chính xác đạt hơn 90%.**

Từ khóa— mạng xã hội, phát hiện tài khoản spam.

I. MỞ ĐẦU

Các mạng xã hội lớn thường hỗ trợ đa ngôn ngữ và cho phép người dùng kết nối với những tài khoản khác trên phạm vi toàn cầu. Khoảng 2 tỷ người dùng Internet đang sử dụng mạng xã hội và con số này được kỳ vọng sẽ tiếp tục tăng lên do xu hướng sử dụng thiết bị di động ngày càng tăng. Theo một thống kê của tạp chí Statista[1], đến tháng Tư năm 2016, Facebook đã trở thành mạng xã hội lớn nhất trên thế giới với hơn 1 tỷ tài khoản đăng ký sử dụng và 1.59 tỷ người dùng hoạt động hàng tháng. Trong những năm gần đây, mạng xã hội trực tuyến ngày càng phụ thuộc vào dữ liệu xã hội để cung cấp những thông tin phù hợp và hữu ích cho người dùng. Trong nhiều nghiên cứu chỉ ra, người dùng Facebook được chỉ hướng tới nội dung dựa trên những gì mà bạn bè và trang họ theo dõi, thích và bình luận[15]. Tuy nhiên, làm sao chúng ta biết được những nội dung từ những tài khoản khác có tin cậy hay không. Từ khi mạng xã hội trở nên phổ biến và quen thuộc trong các hoạt động thường ngày cũng như trong công việc, những tài khoản spam bắt đầu trục lợi từ những người dùng khác bằng những hành vi lừa đảo khả nghi. Hơn nữa, tính mở và sự phụ thuộc tương đối vào người dùng khác cùng với sự phát triển của mạng xã hội cũng góp phần khiến người dùng trở thành đối tượng quan tâm của những tài khoản spam.

[1] http://www.statista.com

Trên Facebook, các trang Page được các tổ chức sử dụng để tương tác với người dùng khác. Người dùng có thể bình luận trên các trang để bạn bè họ có thể biết về sở thích và sự quan tâm của họ, và để nhận được các nội dụng từ trang họ theo dõi trên phần cập nhật trạng thái (News Feed) – một kênh chia sẻ thông tin chính trên Facebook. Có rất nhiều cách để các đối tượng xấu đưa những nội dung spam lên trang như giả mạo tài khoản, phần mềm độc hại, đánh cắp thông tin và tấn công lừa đảo. Tuy vậy, Facebook cũng đã có nhiều các thuật toán phát hiện các toàn khoản giả mạo, nhiều cơ chế chống lừa đảo và các phần mềm độc hại giúp người dùng thật tránh được những nguy cơ lừa đảo. Kết quả là, việc sở hữu nhiều tài khoản giả mạo rất khó và thay vào đó, các đối tượng này sử dụng số ít các tài khoản để thực hiện hành vi thích (Like) hoặc bình luận (Comment) trên nhiều trang Facebook.

Trong nghiên cứu này, chúng tôi tập trung vào bài toán phát hiện tài khoản spam trên Facebook. Các đặc trưng được sử dụng để phát hiện tài khoản spam được dựa trên nội dung và hành vi của người dùng. Chúng tôi đã thu thập tập dữ liệu các bình luận của người dùng trên các trang bán hàng Facebook ở Việt Nam và kết hợp với các hành vi xã hội của người dùng để xây dựng tập dữ liệu của người dùng mạng xã hội. Chúng tôi áp dụng phương pháp Maximum Entropy trên tập dữ liệu để xây dựng mô hình phát hiện tài khoản spam. Thực nghiệm cho kết quả khả quan ban đầu.

Bài báo này gồm các phần như sau. Phần hai trình bày các nghiên cứu liên quan. Mô hình phát hiện tài khoản spam được trình bày chi tiết ở phần ba. Phần bốn bao gồm thực nghiệm và kết quả. Kết luận nghiên cứu được nêu ra ở phần cuối.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Trong phần này, chúng tôi trình bày một số hướng tiếp cận chính của bài toán phát hiện các tài khoản spam trên các mạng xã hội. Với sự phát triển nhanh chóng của các mạng xã hội, vấn đề spam trên mạng xã hội đang thu hút được nhiều sự quan tâm từ cả phía công nghiệp và hàn lâm. Về phía công nghiệp, Facebook đề xuất thuật toán EdgeRank[2], gán cho mỗi bài đăng với một giá trị dựa trên số ít các đặc trưng (ví dụ số lượng thích, số lượng bình luận, số lượng chia sẻ, v.v.). Từ đó, giá trị EdgeRank càng cao thì khả năng là đối tượng spam càng thấp. Nhược điểm của hướng tiếp cận này là các đối tượng spam có thể liên kết với các nhóm của họ và liên tục bình luận và thích các bài đăng của nhau để thu được giá trị EdgeRank cao.

[2] http://techcrunch.com/2010/04/22/facebook-edgerank

Về phía hàn lâm, hầu hết các phương pháp nhận dạng tài khoản spam dựa trên nội dung khai thác được từ các tài khoản cá nhân, các bình luận của họ và các hành vi xã hội của người dùng. Họ sử dụng các thuật toán học máy cho bài toán phân lớp tài khoản spam. Vài năm gần đây, có rất nhiều nghiên cứu trên thế giới về bài toán phát hiện tài khoản spam như [2, 5, 7, 9, 19]. Các nghiên cứu này đưa ra những giải pháp và các dấu hiệu để phân biệt các tài khoản spam trên các mạng xã hội cụ thể, giúp ngăn chặn cũng như loại bỏ các tài khoản đó. Có hai cách tiếp cận phổ biến để xác định tài khoản spam là: phương pháp dựa trên nội dung như [1, 3, 4, 10, 11, 15, 18] và phương pháp đồ thị xã hội ví dụ như [12, 13, 16, 17].

Stringhini và các cộng sự [14] khảo sát các đặc trưng của tài khoản spam bằng cách tạo ra nhiều tài khoản cá nhân trên ba mạng xã hội lớn (Facebook, Twitter và Myspace) và xác định năm đặc trưng có khả năng thường thấy cho việc phát hiện tài khoản giả mạo. Lee và công sự [8] triển khai các nhóm xã hội (social honeypots) bao gồm những tài khoản cá nhân đã xác thực để phát hiện những tài khoản có khả năng là spam, và các máy bot thu thập các dấu hiệu spam bằng cách quét các tài khoản cá nhân của người dùng mà gửi những yêu cầu kết bạn và các đường dẫn liên kết đáng nghi (MySpace và Twitter). Một nghiên cứu khác cũng được thực hiện bởi Alex Hai Wang [15] vào năm 2010, trong đó sử dụng các đặc trưng dựa trên người dùng và nội dung để phát hiện các tài khoản cá nhân spam trên Twitter bằng thuật toán phân lớp Bayes và biểu diễn hành vi người dùng bằng đồ thị xã hội. Những độ đo đánh giá cổ điển được sử dụng để so sánh hiệu suất của các phương pháp phân lớp truyền thống khác nhau như Cây Quyết định, Máy Véc tơ Tựa (SVM), Naïve Bayes và Mạng Nơ ron. Trong số các phương pháp đó, bộ phân lớp Bayes được đánh giá là tốt nhất về hiệu suất. Grier và các cộng sự [6] xác định các đặc trưng liên quan tới nội dung các bài đăng tweet và các đặc trưng cộng đồng trong mạng xã hội Twitter. Các đặc trưng này được sử dụng trong mô hình học máy để phân loại tài khoản cá nhân là spam hay non-spam. Tuy nhiên, những hướng tiếp cận này phụ thuộc vào rất nhiều các đặc trưng mà đòi hỏi khả năng tính toán lớn và nhiều thời gian để xây dựng mô hình huấn luyện. Beutel và các cộng sự [2] đề xuất giải pháp COPYCATCH giúp pháp triển các mô hình Page Like trên Facebook bằng việc phân tích đồ thị xã hội giữa các người dùng, các trang Page và thời gian các cạnh trên đồ thị được tạo ra.

Từ những phân tích trên, chúng tôi xây dựng mô hình phát hiện tài khoản spam dựa trên nội dung và hành vi người dùng. Bằng việc sử dụng phương pháp Maximum Entropy, mô hình xác định người dùng là

tài khoản thật hay là giả mạo và theo các bước sau đây:

- Giải pháp trích chọn đặc trưng được áp dụng để cải thiện kết quả của mô hình thu được.
- Hai tập đặc trưng của dữ liệu được xác định. Thực nghiệm chỉ ra rằng tập đặc trưng là tổ hợp của nội dung bình luận và hành vi người dùng tốt hơn từng tập đặc trưng riêng lẻ.

III. MÔ HÌNH ĐỀ XUẤT

Trong phần này, chúng tôi trình bày mô hình của phương pháp được đề xuất và đưa ra những câu hỏi là động lực cho nghiên cứu của chúng tôi trong mô hình này.

A. Phát biểu bài toán

Bài báo đặt trọng tâm vào việc đề xuất mô hình phát hiện tài khoản spam trên trang Facebook. Trong nghiên cứu này, bài toán được mô tả như sau.

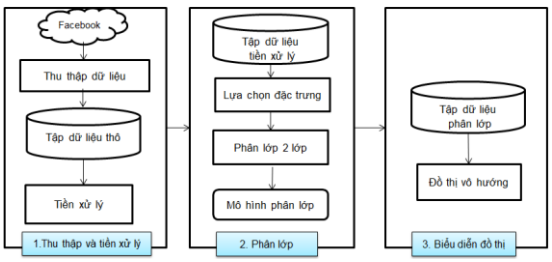
Trong mạng xã hội Facebook, ta có một tập *n* người dùng **U** = {u_1, u_2…u_n} và một tập *m* trang (page) **P** = {p_1, p_2,…p_m }. Mỗi người dùng u_i viết một bình luận trên trang p_j tại thời điểm *t*.

Bài toán phát hiện tài khoản spam là dự đoán liệu u_i có phải là tài khoản spam thông qua bộ phân lớp nhị phân **c**: $u_i \rightarrow$ {spammer, non-spammer}. Để xây dựng **c**, tác giả cần lựa chọn một tập gồm *l* các đặc trưng **F** = {f_1, f_2…f_l} từ tập dữ liệu bình luận của người dùng và hành vi xã hội của người dùng trên trang p_j tại thời điểm *t*.

Trong nghiên cứu này, chúng tôi xem xét hai mô hình dự đoán, có liên quan tới hai cách biểu diễn dữ liệu. Trường hợp thứ nhất, chúng tôi chỉ sử dụng bình luận của người dùng trên trang để xây dựng mô hình dự đoán tài khoản spam. Trường hợp thứ hai, chúng tôi sử dụng cả nội dung bình luận và hành vi của người dùng để xây dựng mô hình dự đoán và so sánh với mô hình đầu tiên.

B. Hướng tiếp cận

Chúng tôi giải quyết bài toán phát hiện tài khoản spam dựa trên nội dung bình luận và hành vi người dùng trên các trang Facebook. **Hình.1** mô tả cấu trúc của mô hình phát hiện tài khoản spam, bao gồm ba pha: Thu thập và tiền xử lý dữ liệu, Phân lớp, Biểu diễn đồ thị.



Hình 1. Mô hình nhận dạng tài khoản spam

1) Pha thu thập và tiền xử lý dữ liệu:

Đầu tiên, chúng tôi dùng API của Facebook và thư viện Restfb để thu thập dữ liệu từ các trang Facebook công khai thông qua tìm kiếm các từ khóa có liên quan, bao gồm thông tin trang, bình luận của người dùng và thông tin người dùng. Mỗi bài viết sẽ chỉ lấy một bài post và tất cả các bình luận trong bài đó. Tiếp theo, chúng tôi trích chọn các thông tin có giá trị như ID người dùng, ID của trang và bình luận; tiền xử lý dữ liệu bằng cách loại đi những từ dừng, các lỗi và tách câu.

Thứ hai, chúng tôi chỉ lấy một phần của tập dữ liệu để gán nhãn, xây dựng tập dữ liệu mẫu. Chúng tôi đọc tất cả các bình luận và gán nhãn cho chúng (hoặc là SPAM hoặc NON-SPAM) dựa trên nội dung của các bình luận. Để xây dựng mô hình phân loại MaxEnt, chúng tôi định nghĩa tập các đặc trưng, bao gồm đặc trưng n-gram và hành vi người dùng. Dựa vào thực nghiệm, chúng tôi thu được kết quả tốt nhất khi kết hợp cả hai đặc trưng 1-gram và 2-gram. Bằng việc khảo sát đặc trưng hành vi người dùng để phát hiện tài khoản spam, chúng tôi tính toán số lượng bình luận của người dùng trong một phút làm đặc trưng cho hành vi người bởi vì một người dùng thật thường sẽ không trả lời 5 bình luận trong 1 phút.

2) Pha phân lớp: Trong pha này, lựa chọn đặc trưng là một phần quan trọng trong Maxent, độ chính xác của mô hình phụ thuộc rất nhiều vào bước này. Chúng tôi lựa chọn các đặc trưng dựa trên bình luận của người dùng và số lượng bình luận của họ trên trang cụ thể. Sau đó, dữ liệu huấn luyện được sử dụng để xây dựng mô hình và kiểm tra, đánh giá kết quả của mô hình. Như được mô tả ở trên, mô hình thứ nhất sử dụng bình luận của người dùng để lựa chọn đặc trưng. Tuy nhiên, chúng tôi bổ sung thêm đặc trưng hành vi người dùng vào đặc trưng dựa trên nội dung bình luận trong mô hình thứ hai. Cuối cùng, việc so sánh kết quả ở hai mô hình này giúp tác giả hiểu rõ hơn về các đặc trưng và sự quan trọng của nó trong mô hình.

Pha phân lớp

3) Pha biểu diễn đồ thị:

Đồ thị mạng song phương với tập các nội dung: $C = \{i\}^N_{i=1}$, trong đó mỗi c_i chứa hai trường thông tin như (nội dung bình luận, ID người dùng) và tập các trang Page: $P = \{j\}^M_{j=1}$. Đồ thị song phương $G = \langle \mathbf{V}, \mathbf{E} \rangle$, tập đỉnh V là tập các trang Page và tập các nội dung (gồm bình luận, ID người dùng và thời gian); tập cạnh E là tập liên kết giữa người dùng với trang Page nếu có bình luận.

E = { (i, j) | liên kết từ người dùng c_j bình luận trên p_j } tại t_{ij}

t_{ij}: là thời gian mà người dùng c_i bình luận trên Page p_j .

IV. THỰC NGHIỆM VÀ ĐÁNH GIÁ

Để đánh gá mô hình trên, chúng tôi xây dựng mô hình thực nghiệm để phát hiện bình luận spam trên các trang Facebook trong lĩnh vực bán hàng. Chúng tôi hiểu rõ các đặc trưng dựa trên nội dung và hành vi người dùng có giá trị trong việc phân biệt tài khoản spam trên miền dữ liệu bán hàng trực tuyến.

A. Dữ liệu thực nghiệm

Chúng tôi thu thập một lượng lớn dữ liệu từ trang Facebook trong cùng một ngày, dựa trên các từ khóa trên các miền. Tập dữ liệu bao gồm 941,038 bình luận bởi 478,496 người dùng từ 23,461 các trang Facebook ở Việt Nam. Chúng tôi lấy một phần của tập dữ liệu để gán nhãn thủ công cho tập dữ liệu mẫu. Dựa trên khảo sát và phân tích trên dữ liệu thu thập được, có một vài dấu hiệu để xác định một bình luận có phải spam hay không như: số lượng người được gắn trong bình luận (tag >=1), có chứa các liên kết (https://; www), và độ dài của bình luận v.v. Dữ liệu đã được gán nhãn được mô tả dưới đây:

BẢNG I. DỮ LIỆU BÌNH LUẬN THEO HAI LỚP

Nhãn	Số lượng
SPAM	**4,864**

Tập dữ liệu mẫu được chia ngẫu nhiên thành 4 tập. Chúng tôi lấy lần lượt ba tập để huấn luyện mô hình và tập còn lại để thực hiện đánh giá chéo 4-fold (4-tập). Phương pháp đánh giá dựa trên dữ liệu gán nhãn có các độ đo thường được sử dụng là độ chính xác (P), độ hồi tưởng (R) và độ đo F1, trong đó đơn vị của các độ đo là % (phần trăm). Kết quả thực nghiệm sẽ được trình bày ở phần B.

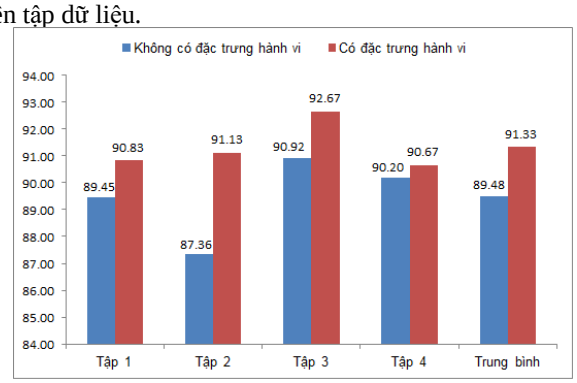
B. Kết quả thực nghiệm và phân tích

BẢNG II. KẾT QUẢ ĐỘ ĐO LẦN KIỂM TRA TỐT NHẤT

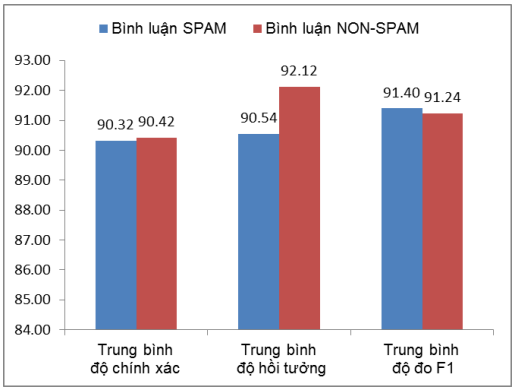
Lớp	Gán nhãn	Dự đoán	TP	P	R	F1
SPAM	1197	1164	1093	93.90	91.31	92.59
NON-SPAM	1192	1225	1121	91.51	94.04	92.76
$F1_{macro}$						92.69
$F1_{micro}$						92.67

Bảng II chỉ ra kết quả thực nghiệm của lần kiểm tra thứ ba, trong đó TP là bình luận được gán nhãn đúng so với mô hình. Chúng tôi thu được kết quả cao khi sử dụng cả hai đặc trưng dựa trên nội dụng và đặc trưng hành vi người dùng.

Hình.2 cho thấy độ đo F1 của cả bốn tập con và giá trị trung bình trên bốn tập con. Với mỗi tập con, lần đầu tiên là thí nghiệm chỉ sử dụng đặc trưng dựa trên nội dung, trong khi lần thứ hai sử dụng cả đặc trưng dựa trên nội dung lẫn hành vi người dung. Như chúng ta thấy, bộ phận lớp sử dụng thêm đặc trưng hành vi cho kết quả tốt hơn. Đặc trưng hành vi người dùng có thể cải thiện độ đo F_1 khoảng hơn 2%. Từ kết quả của kiểm tra đánh giá chéo 4-fold, kết quả thu được khá ổn định trên cả bốn tập con. Điều này cho thấy rằng mô hình phân loại hoạt động hiệu quả trên tập dữ liệu.



Hình 2. Kết quả độ đo F_1 của kiểm thử chéo 4 tập



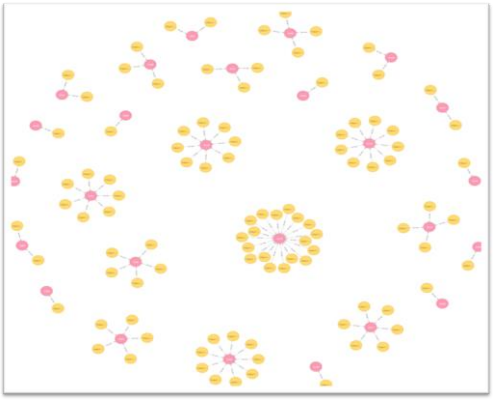
Hình 3. Kết quả trung bình của độ chính xác, độ hồi tưởng và giá trị F_1 của nhãn Spam and Non-Spam trên 4 tập con (với đặc trưng Hành vi người dùng)

Từ kết quả, chúng tôi nhận thấy những dấu hiệu spam rất hữu ích trong quá trình gán nhãn xây dựng tập dữ liệu mẫu. Đó là những đặc trưng có thể phân biệt được giữa các bình luận spam và không spam, ví dụ như số lượng người được tag (≥1), đường liên kết, số lượng các kí tự (≤5 hay ≥100), v.v. Thực tế,

những người dùng bình luận quá ngắn hoặc quá dài sẽ có nhiều khả năng là các tài khoản spam.

Chúng tôi cũng đánh giá trung bình của độ chính xác, độ hồi tưởng và độ đo F1 của hai lớp: SPAM và NON-SPAM trên 4 tập dữ liệu con. Kết quả được trình bày trong hình 3. Kết quả của lớp NON-SPAM xấp xỉ bằng kết quả của lớp SPAM. Điều này một phần do số lượng các bình luận mang nhãn NON-SPAM gần bằng các bình luận với nhãn SPAM (4,692 với 4,864).

Đồ thị được biểu diễn bằng Neo4j, so sánh không gian phụ thuộc vào số lượng page và khoảng thời gian xét từ bình luận đầu tiên tới bình luận cuối cùng trên bài viết trong một ngày làm mốc để xét. Với không gian nhỏ, số lượng Page bán hàng < 50 Page, bình luận spam < 100, còn không gian lớn > 50 Page, bình luận spam > 1000.



Hình 4. Kết quả biểu diễn đồ thị

Kết quả biểu diễn đồ thị cho không gian nhỏ và không gian lớn có một điểm chung là: có những page có số lượng bình luận spam bùng nổ (> 20) so với những page chỉ có một vài bình luận spam. Điều này có thể chứng tỏ rằng những những tài khoản spam có hành vi bình luận spam cùng nhau và những bài viết có số bình luận spam nhỏ hơn 5 thì có thể coi là không phải cuộc tấn công spam. Những bình luận spam cùng lúc với số lượng lớn nhằm mục đích quảng cáo sản phẩm của các trang Page nhằm gây sự chú ý của người dùng trên mạng xã hội Facebook.

Tuy vậy, mô hình phân loại hoạt động không tốt với một vài loại bình luận. Một vài bình luận chỉ gồm một người được gắn thẻ (tag) và một vài kí tự viết tắt sẽ gây khó khăn trong quá trình phân loại.

Tài khoản spam thường gắn các người dùng khác vào bình luận để quảng cáo hoặc thu hút sự chú ý trong khi những tài khoản xác thực chỉ tag bạn bè của họ với một vài kí tự ngắn. Khảo sát trên tập dữ liệu bình luận chỉ ra rằng việc tag bạn bè với đoạn kí tự quá dài hoặc không có kí tự kèm theo sẽ được gán nhãn là spam, trong khi đó việc tag người dùng khác với một vài thông tin bổ sung như địa chỉ và số điện thoại sẽ được gán nhãn là không spam. Hơn nữa, chúng tôi cũng gặp khó khăn trong việc gán nhãn các bình luận chứa các tình cảm hay cảm xúc như: lời khen, sự chê bai v.v. Để giải quyết các trường hợp này, tác giả cần kết hợp thêm nhiều đặc trưng khác cho tới việc xử lý cú pháp v.v.

V. KẾT LUẬN

Như vậy, trong bài báo này chúng tôi đã xây dựng mô hình phân lớp dựa trên phương pháp Maximum Entropy để phân loại bình luận người dùng từ các trang Facebook ở Việt Nam vào nhãn SPAM hoặc NON-SPAM. Việc kết hợp những đặc trưng dựa trên nội dung và hành vi người dùng đã góp phần đáng kể để thu được kết quả tốt nhất. Qua thực nghiệm, chúng tôi đã đạt được giá trị trung bình F1 hơn 90%, một kết quả rất khả quan cho những nghiên cứu tiếp theo của bài toán này. Kết quả cũng chỉ ra hướng tiếp cận đúng đắn của chúng tôi trong việc phát hiện các bình luận spam bằng việc sử dụng các dấu hiệu phù hợp. Chúng tôi cũng thấy rằng cần phải thêm những đặc trưng tốt hơn cũng như cải thiện được chất lượng của tập dữ liệu mẫu của mô hình để có thể phân biệt hiệu quả các bình luận còn nhập nhằng. Đây cũng là trọng tâm nghiên cứu tiếp theo của tác giả trong tương lai.

TÀI LIỆU THAM KHẢO

[1] F. Benevenuto, T. Rodrigues, V. A. F. Almeida, J. M. Almeida, and M. A. Gonçalves, “Detecting spammers and content promoters in online video social networks,” in SIGIR 2009, pp. 620-627.

[2] A. Beutel, W. Xu, V. Guruswami, C. Palow, and C. Faloutsos, “CopyCatch: stopping group attacks by spotting lockstep behavior in social networks,” WWW 2013, pp. 119-130.

[3] M. Fazeen, R. Dantu, and P. Guturu, “Identification of leaders, lurkers, associates and spammers in a social network: context-dependent and context-independent approaches,” Social Netw. Analys. Mining, vol. 1, no. 3, pp. 241-254, 2011.

[4] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao, “Detecting and characterizing social spam campaigns,” ACM Conference on Computer and Communications Security, pp. 681-683, 2010.

[5] M. K. Girish Khurana, “Review: Efficient Spam Detection on Social Network,” ISSN, vol. 3, no. 6, pp. 2321-9653, 2015.

[6] C. Grier, K. Thomas, V. Paxson, and C. M. Zhang, “@spam: the underground on 140 characters or less,” in ACM Conference on

Computer and Communications Security, 2010, pp. 27-37.

[7] K. Lee, J. Caverlee, and S. Webb, "The social honeypot project: protecting online communities from spammers," in WWW 2010, pp. 1139-1140.

[8] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: social honeypots + machine learning," in SIGIR 2010, pp. 435-442.

[9] L. Liu, Y. Lu, Y. Luo, R. Zhang, L. Itti, and J. Lu, "Detecting "Smart" Spammers On Social Network: A Topic Model Approach," CoRR, vol. abs/1604.08504, 2016.

[10] M. McCord, and M. Chuah, "Spam Detection on Twitter Using Traditional Classifiers," in ATC 2011, pp. 175-186.

[11] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang, "Twitter spammer detection using data stream clustering," Inf. Sci., vol. 260, pp. 64-73, 2014.

[12] S. Rayana, and L. Akoglu, "Collective Opinion Spam Detection: Bridging Review Networks and Metadata," in KDD 2015, pp. 985-994.

[13] B. Stone-Gross, T. Holz, G. Stringhini, and G. Vigna, "The Underground Economy of Spam: A Botmaster's Perspective of Coordinating Large-Scale Spam Campaigns," in LEET 2011.

[14] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in ACSAC 2010, pp. 1-9.

[15] A. H. Wang, "Detecting Spam Bots in Online Social Networking Sites: A Machine Learning Approach," in DBSec, 2010, pp. 335-342.

[16] A. H. Wang, "Don't Follow Me - Spam Detection in Twitter," in SECRYPT 2010, pp. 142-151.

[17] C. Wilson, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao, "Beyond Social Graphs: User Interactions in Online Social Networks and their Implications," TWEB, vol. 6, no. 4, pp. 17, 2012.

[18] S. Yardi, D. M. Romero, G. Schoenebeck, and D. Boyd, "Detecting Spam in a Twitter Network," First Monday, vol. 15, no. 1, 2010.

[19] X. Zhang, and X. Zheng, "A novel method for spammer detection in social networks," in ICSDM 2015, pp. 115-118.