Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

PHÂN LOẠI Ý KIẾN TRÊN TWITTER

Võ Tuyết Ngân[1] và Đỗ Thanh Nghị[2]

[1] *Khoa chuyên ngành, Trường Cao đẳng Cộng đồng Cà Mau*
[2] *Khoa Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ*

Thông tin chung:
Ngày nhận: 19/09/2015
Ngày chấp nhận: 10/10/2015

Title:
Twitter sentiment analysis

Từ khóa:
Phân loại văn bản, phân loại ý kiến, mô hình túi từ Bow, máy học vectơ hỗ trợ SVM, giải thuật Naïve Bayes, mạng ngữ nghĩa

Keywords:
Text classification, Twitter sentiment analysis, Bag-of-Words-(Bow), Support Vector Machines (SVM), Multinomial Naïve Bayes (MNB), WordNet

ABSTRACT

Twitter sentiment analysis aims at classifying the comment into positive or negative sentiment. In this paper, we propose to use the bag-of-words model and the Multinomial Naïve Bayes algorithm for dealing with the sentiment classification task. In the first step, raw data sets are the comments on Twitter collected following topic. It is necessary to perform the preprocessing task, including the special characters of Twitter, continuously repeatable characters, acronyms, slang, emoticons, WordNet, and representation in Bow model. Preprocessing stage provides the large dimensional datasets in which almost values (about 99%) are zero. And then, the data set is stored in the LibSVM format (dim_index: non_zero_value). This strategy is to reduce the memory complexity and also require our new implementation of Multinomial Naïve Bayes (MNB) for dealing with the new data format. The experimental results on the data sets show that our implementation of Multinomial Naïve Bayes (MNB) algorithm is very simple and accurate.

TÓM TẮT

Phân loại ý kiến trên Twitter là phân loại cho từng bình luận theo hướng quan điểm tích cực hay tiêu cực dựa trên nội dung bình luận. Trong bài viết này, chúng tôi đề xuất sử dụng mô hình túi từ và giải thuật máy học Multinomial Naïve Bayes để phân loại ý kiến. Ở bước đầu tiên, từ tập dữ liệu thô là những ý kiến trên Twitter được thu thập theo chủ đề, chúng tôi tiến hành tiền xử lý các kí tự đặc biệt của Twitter, các kí tự trùng lặp gần nhau, từ viết tắt, tiếng lóng, biểu tượng cảm xúc, mạng ngữ nghĩa, biểu diễn văn bản theo mô hình túi từ. Giai đoạn tiền xử lý cho ra tập dữ liệu có số chiều lớn, nhưng trong đó đa số (khoảng 99%) các giá trị bằng 0. Để tiết kiệm bộ nhớ, chiến lược lưu trữ chỉ lưu những giá trị khác 0 (theo định dạng LibSVM). Cách lưu trữ này dẫn đến yêu cầu cài đặt lại giải thuật máy học Multinomial Naïve Bayes để có thể xử lý định dạng mới của tập dữ liệu. Kết quả thực nghiệm trên các tập dữ liệu cho thấy bản cài đặt mới của giải thuật Multinomial Naïve Bayes (MNB) phân lớp hiệu quả, đơn giản và chính xác.

1 GIỚI THIỆU

Hiện nay, công nghệ ngày càng phát triển, đặc biệt với sự ra đời của mạng xã hội, lượng thông tin trên mạng xã hội là một kho dữ liệu lớn, có nhiều tri thức hữu dụng nhưng tiềm ẩn. Vấn đề quan trọng hiện nay là làm thế nào để khai thác kho dữ liệu khổng lồ này. Dữ liệu trên mạng xã hội thường

lớn, phi cấu trúc, phức tạp, thậm chí là các thông tin rác cũng rất nhiều. Cần thiết phải có những nghiên cứu để xác định được thông tin gì là cần thiết và thông tin nào là dư thừa. Các nhà nghiên cứu xử lý ngôn ngữ tự nhiên và trích chọn thông tin đều đi tìm câu trả lời cho câu hỏi đó. Môi trường của mạng xã hội khá tự do, nơi cảm xúc cá nhân được đề cao, là nơi có thể dễ dàng thu thập những ý kiến của người dùng về một sản phẩm nào đó. Mỗi người quan tâm và đánh giá về một số lĩnh vực, mỗi người cần biết thị hiếu, xu hướng về một vài thứ mà trên mạng xã hội thì bao gồm vô số thông tin về các lĩnh vực mà đa số người đều quan tâm. Các doanh nghiệp, khách hàng thường quan tâm đến sản phẩm. Người dùng thì quan tâm sản phẩm này có tốt không, sản phẩm kia tốt ở chỗ nào và chỗ nào không tốt. Còn doanh nghiệp thì lại thường quan tâm đến sản phẩm của họ được mọi người tiếp đón thế nào, những điểm nào chưa tốt để họ bổ sung sửa chữa, hay phát huy thêm những điểm người dùng quan tâm. Câu trả lời cho những câu hỏi này nằm trong nghiên cứu về "Opinion Mining" hay còn gọi "phân tích ý kiến người dùng".

Thông thường, để đánh giá về một sản phẩm nào đó, nhà nghiên cứu sẽ trích chọn những đặc điểm riêng (Features) của sản phẩm. Sau đó từ những review, comment, Feedback,… đánh giá xem tính năng của sản phẩm này được mọi người tiếp đón thế nào (Huifeng Tang *et al*., 2009).

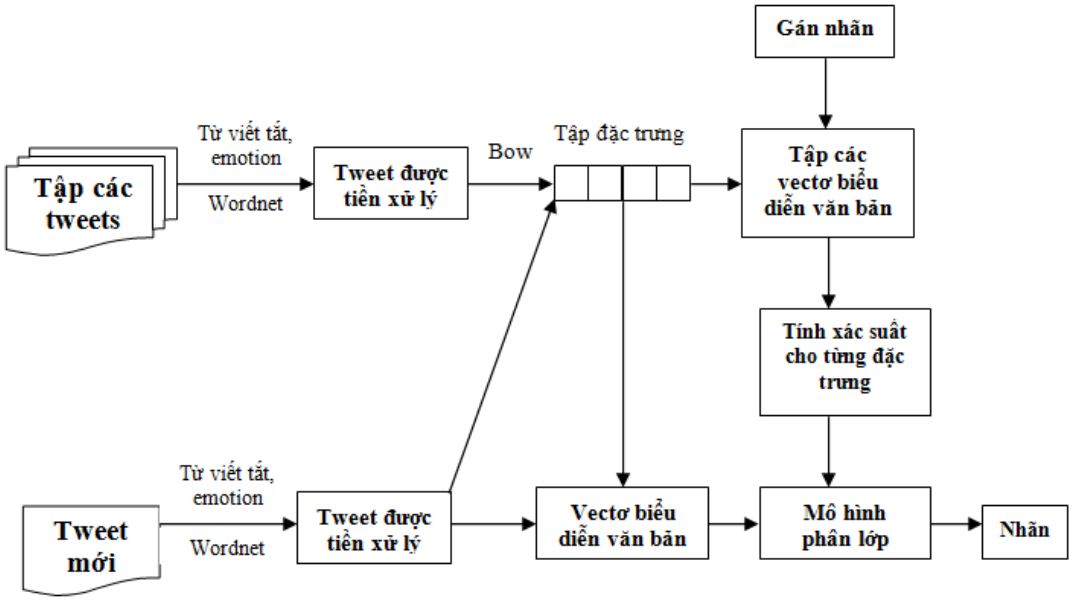
Vài năm trở lại đây, phân loại ý kiến trên mạng xã hội Twitter là chủ đề nóng giữa các nhà nghiên cứu. Ưu điểm của Twitter là cho phép người dùng tìm kiếm những bình luận (tweet) theo từ khóa của họ, hơn nữa Twitter có hỗ trợ các API cho phép sao chép những bình luận này về, tạo thuận lợi cho việc nghiên cứu. Tuy nhiên, mỗi bình luận của Twitter chỉ giới hạn 140 kí tự, lượng thông tin thu về khá hạn chế nên việc phân tích ý kiến gặp phải nhiều khó khăn.

Cho tới nay, phần lớn nghiên cứu phân loại ý kiến trên Twitter tập trung vào các đặc trưng, họ lựa chọn những đặc trưng riêng cho từng sản phẩm, đa số là dựa vào sự lựa chọn từ loại và kết hợp từ loại: danh từ, động từ, tính từ và trạng từ, đặc biệt là 2 loại từ cuối cùng. Độ chính xác khá cao trên 80% (Go *et al*., 2008). Tuy nhiên, điều này mang tính chủ quan vì các loại từ khác cũng có thể mang lại tính hiệu quả trong việc phân loại ý kiến.

Một nhóm nghiên cứu khác (Turney, 2002) định nghĩa các biểu hiện quan điểm trong thuật ngữ "ý kiến" của chính họ dựa trên biểu thị tính đối lập, POS tagging và loại chủ đề (sản phẩm) đối với các từ chỉ ý kiến tương ứng. Tuy nhiên, phương pháp này lại không được sử dụng rộng rãi và kết quả mang lại chưa cao. Cũng do một thực tế là các thuật ngữ ý kiến độc lập này không có giá trị công khai, các nhóm phân tích ý kiến không thể cộng tác với nhau và như vậy thì không thể đưa ra một tài liệu tiêu chuẩn cho việc sử dụng trong tương lai.

Theo kết quả nghiên cứu, các nhà nghiên cứu cho rằng sử dụng mô hình unigram kết hợp với giải thuật máy học Multinomial Naïve Bayes (MNB) đem lại hiệu quả cao, hơn nữa việc cài đặt khá đơn giản và mang tính khách quan. (Bermingham & Smeaton, 2010) đã thu thập các ý kiến trên các blog, microblog (Twitter) để phân loại. Kết quả đạt được độ chính xác 74.85% cho phân lớp nhị phân (tích cực, tiêu cực) dựa trên mô hình Unigram đối với microblog. Ngoài ra, hai ông còn đưa ra kết luận là độ chính xác của phân loại ý kiến trên microblog cao hơn blog, mặc dù lượng thông tin trong các ý kiến của blog nhiều hơn, vectơ đặc trưng không thưa như microblog. Đối với những ý kiến ngắn, MNB cho kết quả tốt hơn SVM

Xuất phát từ nhu cầu thực tiễn trên, chúng tôi xin đề xuất cài đặt giải thuật MNB, mô hình unigram để phân loại ý kiến của Twitter. Nghiên cứu sử dụng bộ từ vựng trên 15000 từ bao gồm nhiều từ loại, đề xuất cách lưu trữ dữ liệu thưa để tiết kiệm bộ nhớ đối với số lượng bình luận (tweet) lớn, hay gặp phải nếu chủ đề cần tìm được nhiều người quan tâm.



Hình 1: Sơ đồ phân lớp ý kiến với giải thuật MNB

2 PHƯƠNG PHÁP NGHIÊN CỨU

2.1 Tiền xử lý dữ liệu

Dữ liệu của mạng Twitter rất phức tạp, phi cấu trúc, nhiễu. Để có thể phân lớp bằng giải thuật máy học, trước hết cần phải thực hiện các thao tác tiền xử lý.

Các tên tài khoản Twitter của người dùng được chuyển về dạng USER_AT.

Địa chỉ trang web được chuyển về dạng URL.

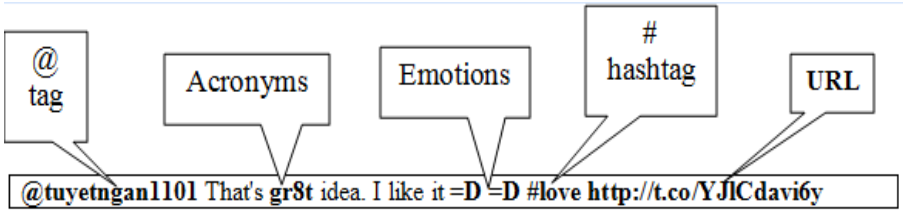
Chúng tôi còn xử lý thêm các ký tự trùng lắp gần nhau, biểu tượng cảm xúc, từ viết tắt, tiếng lóng, mạng ngữ nghĩa.

Xử lý các kí tự trùng lắp gần nhau: Để loại bỏ ký tự trùng lắp gần nhau, đầu tiên phải tìm ra dãy chứa các kí tự trùng lắp, sau đó tiến hành loại bỏ các kí tự trùng lắp dựa trên biểu thức chính quy. Do trong tiếng Anh, các nguyên âm giống nhau thường đứng cạnh nhau nên việc loại bỏ gặp khá nhiều khó khăn. Ở đây chúng tôi sẽ lấy 2 kí tự nằm cạnh nhau trong dãy các tự trùng lắp gần nhau.

Xử lý biểu tượng cảm xúc: Chúng tôi sử dụng những emotion phổ biến của yahoo và facebook (gồm 36 biểu tượng cảm xúc), vì đây là những emotion được sử dụng lâu đời và phổ biến. Với mỗi tweet có chứa tổ hợp kí tự đại diện cho biểu tượng cảm xúc, lần lượt thay thế các kí tự đó bằng ý nghĩa đã được lưu trong bộ từ điển.

Xử lý từ viết tắt và tiếng lóng: Bất kì mạng xã hội nào cũng chứa một vốn từ vựng là từ viết tắt và những tiếng lóng do người sử dụng tạo ra. Twitter cũng vậy, điều này gây ra nhập nhằng trong việc tiền xử lý ngôn ngữ. Trong đề tài nghiên cứu này, chúng tôi sử dụng hơn 100 thuật ngữ của Twitter bao gồm: chữ viết tắt cũng các từ, cụm từ, công cụ kĩ thuật có liên quan đến Twitter, được người dùng mạng xã hội này quy ước, sưu tập và truyền bá. Tương tự như phương pháp xử lý biểu tượng cảm xúc, lần lượt thay thế các từ viết tắt bằng ý nghĩa của chúng trong mỗi tweet.

Xử lý mạng ngữ nghĩa: Mô hình WordNet là một loại từ điển tương tự từ điển đồng nghĩa. WordNet phân chia từ vựng thành 5 loại: noun, verb, adjective, adverb và funtion words, nhưng thực tế nó chỉ chứa noun, verb, adjective, adverb. Những tập đồng nghĩa được gọi là synset (SYN onym SET = synset) tự thân chúng không giải thích về nghĩa (hay ý niệm) mà chúng mang là gì, chúng chỉ cho biết là chúng có mang một nghĩa (ý niệm) duy nhất nào đó mà tất cả các từ có dạng từ được chứa trong tập đó cùng mang. Trong nghiên cứu này, chúng tôi sử dụng phương pháp xây dựng SentiWordNet dựa trên việc phân tích số lượng những lời nhận xét được kết nối với các synset, và dựa trên kết quả về vectơ đại diện cho phân lớp synset bán giám sát.



Xử lý tag, hashtag, URL:

AT_USER That's **gr8t** idea. I like it **=D =D** *love URL*

Xử lý biểu tượng cảm xúc:

AT_USER That's **gr8t** idea. I like it *happy happy love URL*

Xử lý từ viết tắt:

AT_USER That's *great* idea. I like it *happy happy love URL*

Xử lý wordnet:

AT_USER That's *great* idea. I like it *happy happy love URL outstanding bang-up bully corking cracking dandy groovy keen neat nifty not bad peachy slap-up swell smashing capital majuscule big enceinte expectant gravid great large heavy with child iodine iodin atomic number 53 ace single unity I one i ane the like the likes of ilk wish care same alike comparable corresponding like information technology IT felicitous glad happy well-chosen felicitous glad happy well-chosen passion beloved dear dearest honey erotic love sexual love lovemaking making love love life enjoy sleep together roll in the hay love make out make love sleep with get laid have sex know do it be intimate have intercourse have it away have it off screw fuck jazz eff hump lie with bed have a go at it bang get it on bonk URL uniform resource locator universal resource locator*

Hình 2: Ví dụ về các bước tiền xử lý dữ liệu

2.2 Biểu diễn dữ liệu

Mô hình Bag of Words (BoW) là một mô hình được sử dụng phổ biến trong lĩnh vực phân loại văn bản. Mô hình này thường sử dụng để xử lý ngôn ngữ tự nhiên, được dùng để biểu diễn tài liệu, xem tài liệu là một tập hợp các từ (words) mà không quan tâm đến thứ tự cũng như cấu trúc cú pháp của chúng.

Một văn bản được biểu diễn dạng véc-tơ (có *n* thành phần là các từ tương ứng) mà giá trị thành phần thứ *j* là tần số xuất hiện từ thứ *j* trong văn bản. Nếu xét tập *D* gồm *m* văn bản và tự điển có *n* từ vựng, thì *D* có thể được biểu diễn thành bảng *D* kích thước *m x n*, dòng thứ *i* của bảng là véc-tơ biểu diễn văn bản thứ *i* tương ứng.

Giả sử dữ liệu có 15.000 tweets với 20.000 đặc trưng (từ vựng), thông thường mỗi tweet sẽ được lưu trữ như sau:

Chỉ mục	1	2	3	…	20000
Tần số xuất hiện	0	1	0	…	0

Nếu mỗi tần số xuất hiện của 1 đặc trưng tốn khoảng 2 bytes để lưu trữ, vậy mỗi tweet tốn 40.000 bytes. 15.000 tweets tốn 600.000.000 bytes.

Đối với nghiên cứu này, chúng tôi đề xuất cách lưu trữ tiết kiệm bộ nhớ, tương tự như LibSVM [Chang & Lin, 2011], chỉ lưu những từ có tần số xuất hiện lớn hơn 0. Cách lưu trữ như sau:

<label> <index-1>:<value-1> <index-2>:<value-2> ...

Trong đó:

<label> là lớp ban đầu của tweet, 1 là tích cực, 0 là tiêu cực.

<index-i> chỉ mục của từ thứ i.

<value-i> tần số xuất hiện của từ i.

Vì mỗi bình luận trên Twitter chỉ giới hạn 140 kí tự, nên số lượng các từ xuất hiện rất ít, trung bình từ 5 – 7 từ khi chưa xử lý wordnet, 10 - 12 từ khi đã xử lý wordnet. Nếu phải lưu tất cả các tần số xuất hiện của từng từ trong tweets, dữ liệu sẽ trở nên rất thưa, đa số đều mang giá trị 0, dẫn đến sự lãng phí bộ nhớ.

Nếu lưu trữ tiết kiệm bộ nhớ, trung bình sẽ có 6 đặc trưng tần số xuất hiện lớn hơn 0:

2:1 5:1 101:1 609:1 1200:1 15356:1

Với mỗi đặc trưng có dạng **index-i:value-i**, ta tốn khoảng 5 bytes, vậy 1 tweet chỉ tốn trung bình 30 bytes. 15.000 tweet sẽ chiếm 450.000 bytes. So với cách lưu trữ ban đầu, chúng ta tiết kiệm được 599.550.000 bytes.

Tất nhiên 15.000 tweets chỉ là một con số vô cùng nhỏ so với lượng dữ liệu trên Twitter. Nếu dữ liệu càng lớn, ý nghĩa của việc lưu trữ tiết kiệm bộ nhớ sẽ được thể hiện càng rõ.

2.3 Phân loại ý kiến bằng giải thuật máy học MNB

Multinomial Naïve Bayes (MNB) là một mô hình đơn giản nhưng hoạt động rất tốt trong việc phân loại văn bản. [Lewis & Gale, 1994] đã đề xuất kết hợp mô hình túi từ và NB tạo ra giải thuật Multinomial Naïve Bayes. Cụ thể trong bài toán của chúng ta như sau:

Gọi C là tập hợp các lớp của văn bản (C có 2 phần tử +1 và -1). Gọi t_i là một văn bản mới đến. Ta chọn xác suất để t_i thuộc vào lớp c_i lớn nhất. Xác suất này được tính bởi công thức:

$$\Pr(c \mid t_i) = \frac{\Pr(c).\Pr(t_i \mid c)}{\Pr(t_i)}$$

Với $c \in C$

Chú ý:

Pr(c) được tính bằng tổng số văn bản của lớp *c* chia cho tổng số văn bản của tất cả các lớp.

Khi tìm giá trị lớn nhất của $Pr(c|t_i)$ ta có thể bỏ qua tính $Pr(t_i)$ do không đổi khi so sánh.

Xác suất $Pr(t_i|c)$ được tính bằng công thức:

$$\Pr(t_i \mid c) = (\sum_n f_{ni})! \prod_n \frac{\Pr(w_n \mid c)^{f_{ni}}}{f_{ni}!}$$

Chú ý:

f_{ni} là tần suất từ thứ *n* trong t_i.

$Pr(w_n|c)$ là xác suất của từ thứ *n* khi cho trước lớp *c*.

Thay $(\sum f_{ni})!, \prod f_{ni}!$ là α, ta có công thức $\Pr(t_i \mid c) = \alpha \prod_n \Pr(w_n \mid c)^{f_{ni}}$

3 KẾT QUẢ VÀ THẢO LUẬN

Để đánh giá hiệu quả của phương pháp đề xuất, chúng tôi đã thực hiện cài đặt giải thuật MNB (Lewis & Gale, 1994) (mô đun phân loại ý kiến trên Twitter), sử dụng ngôn ngữ Python và thư viện wordnet NLTK của nó, đồng thời chúng tôi đã thay đổi cấu trúc chương trình thích hợp với cách lưu trữ tiết kiệm bộ nhớ. Chúng tôi sử dụng mô đun biểu diễn dữ liệu theo mô hình túi từ BoW (McCallum, 1988). Ngoài ra, chúng tôi cũng cần so sánh MNB với một giải thuật SVM chuẩn, được sử dụng phổ biến trong cộng đồng máy học là LibSVM (Chang & Lin, 2011).

Về dữ liệu thực nghiệm, chúng tôi sử dụng tập dữ liệu được sưu tập bởi [Go *et al*., 2009] được lấy từ các API thu thập theo định kì trên Twitter. Các tweets được chép trong khoảng thời gian từ ngày 06/04/2009 đến ngày 25/6/2009 với 72 chủ đề thuộc nhiều lĩnh vực: mua bán, kĩ thuật, âm nhạc, khu vực,… Kết quả ông thu được 1 triệu 6 tweets với 8000 bình luận tích cực và 8000 bình luận tiêu cực không trùng nhau.

Bộ dữ liệu 1 (bộ dữ liệu gốc): 15.000 bình luận được lấy ngẫu nhiên trong bộ dữ liệu 1 triệu 6 của (Go *et al*., 2009).

Bộ dữ liệu 2: là bộ dữ liệu gốc được xử lý biểu tượng cảm xúc.

Bộ dữ liệu 3: là bộ dữ liệu 2 được xử lý từ viết tắt.

Bộ dữ liệu 4: là bộ dữ liệu 3 được xử lý mạng ngữ nghĩa.

Chúng tôi sử dụng nghi thức kiểm tra hold – out để đánh giá hiệu quả của 2 giải thuật phân lớp, lấy ngẫu nhiên 2/3 tập dữ liệu để học (10000 bình luận) và 1/3 tập dữ liệu kiểm tra (5000 bình luận), thực hiện trên cùng một tập dữ liệu mẫu.

Chúng tôi tiến hành so sánh kết quả dựa trên các tiêu chí như:

TP: tổng số phần tử của lớp tích cực được mô hình phân lớp là tích cực.

TN: tổng số phần tử của lớp tiêu cực được mô hình phân lớp là tiêu cực.

FN: tổng số phần tử của lớp tích cực bị mô hình phân lớp sai thành tiêu cực.

FP: tổng số phần tử của lớp tiêu cực bị mô hình phân lớp sai thành tích cực.

TP rate (*recall*): độ chính xác của lớp tích cực.

TN rate (*recall*): độ chính xác của lớp tiêu cực.

Precision: là số phần tử được mô hình phân lớp đúng về lớp tích cực chia cho tổng số phần tử được dự đoán là lớp tích cực.

Accuracy: độ chính xác toàn cục.

F1: trung bình điều hòa của precision và recall.

Sau khi tiến hành thử nghiệm nhiều lần giải thuật SVM, chúng tôi nhận thấy rằng sử dụng hàm nhân tuyến tính với hằng số cost = 1 (dung hòa độ lớn của lề phân hoạch và lỗi) cho kết quả phân lớp cao nhất. Tất cả các giải thuật trên đều được thực hiện trên máy tính cá nhân (Intel Pentium T3400, 2.2 GHz, 2GB RAM] chạy hệ điều hành ubuntu 12.04

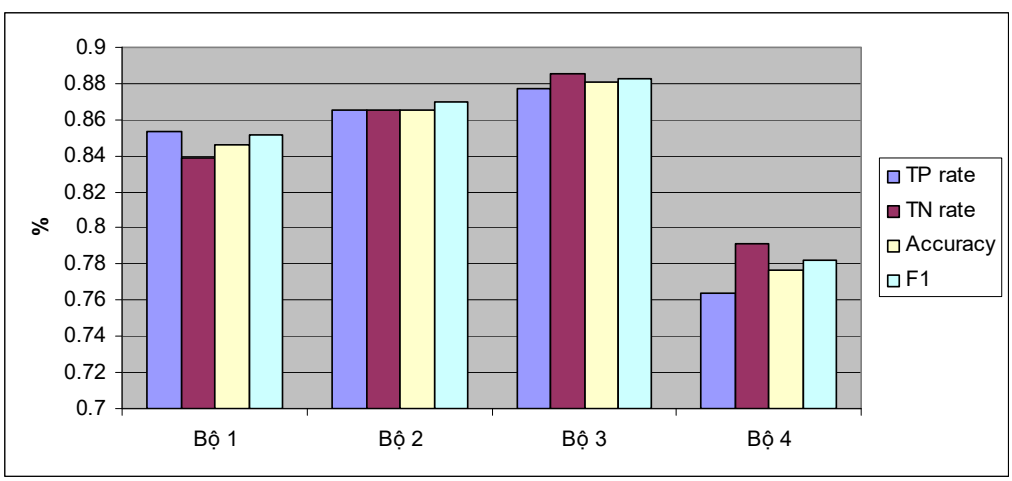
Sau khi chúng tôi tiến hành phân lớp trên 4 bộ dữ liệu, kết quả thu được từ 2 giải thuật máy học như trình bày trong Bảng 1, biểu đồ trong Hình 3.

Dựa vào biểu đồ Hình 3, ta thấy giải thuật MNB phân lớp chính xác khi so sánh với giải thuật SVM. MNB có thể lưu trữ tiết kiệm bộ nhớ theo định dạng của LibSVM, hơn nữa độ chính xác tổng thể của MNB cao hơn gần 10% so với SVM.

Sau 2 bước tiền xử lý, hiệu quả của giải thuật tăng lên, tuy nhiên mức độ tăng chậm. Riêng đối với dữ liệu xử lý WordNet, các chỉ số giảm xuống đáng kể, kết quả thấp hơn khi chưa xử lý. Ngoài ra, tỉ lệ số phần tử lớp tích cực được dự đoán là lớp tích cực cao hơn tỉ lệ số phần tử lớp tiêu cực được dự đoán là lớp tiêu cực, ngoại trừ dữ liệu được xử lý WordNet.

Bảng 1: Kết quả phân lớp ý kiến bằng 2 giải thuật MNB và SVM

	MNB				SVM			
	Bộ 1	**Bộ 2**	**Bộ 3**	**Bộ 4**	**Bộ 1**	**Bộ 2**	**Bộ 3**	**Bộ 4**
TP	2200	2248	2228	2000	1836	1921	1889	1703
FN	377	350	312	619	741	619	709	811
TN	2032	2079	2177	1884	1674	1616	1665	1662
FP	391	323	283	497	749	844	737	824
TP rate = Recall	0.853706	0.865281	0.877165	0.76365	0.712456	0.756299	0.727098	0.677407
TN rate	0.83863	0.865529	0.884959	0.791264	0.690879	0.656911	0.693172	0.668544
Precision	0.849093	0.874368	0.887296	0.800961	0.710251	0.694756	0.719345	0.673922
Accuracy	0.8464	0.8654	0.881	0.7768	0.702	0.7074	0.7108	0.673
F1	0.851393	0.869801	0.882202	0.781861	0.711352	0.724222	0.723201	0.67566



Hình 3: Biểu đồ so sánh kết quả của 4 bộ dữ liệu bằng giải thuật MNB

Kết quả thực nghiệm cho phép chúng tôi tin rằng giải thuật MNB phân lớp ý kiến trên Twitter hiệu quả, kể cả số chiều lớn.

4 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Chúng tôi vừa trình bày một hướng tiếp cận trong việc phân loại ý kiến mạng xã hội, sử dụng phương pháp biểu diễn văn bản bằng mô hình túi từ và giải thuật máy học MNB. Mô hình túi từ được xây dựng đơn giản, nhanh, để biểu diễn văn bản dưới dạng véc-tơ tần số xuất hiện của từ trong văn bản, số chiều lớn. Thay vì lưu trữ đầy đủ giá trị của vectơ, chúng tôi đề xuất cách lưu trữ theo chuẩn LibSVM để tiết kiệm bộ nhớ. Chúng tôi đã cài đặt lại giải thuật máy học Multinomial Naïve Bayes để có thể xử lý định dạng mới của tập dữ liệu. Kết quả thực nghiệm trên các tập dữ liệu cho thấy bản cài đặt mới của giải thuật Multinomial Naïve Bayes (MNB) phân lớp hiệu quả, đơn giản và chính xác khi so sánh với máy học SVM.

Trong tương lai, chúng tôi tiếp tục nghiên cứu cho những chủ đề nhất định và các mạng xã hội khác, đặc biệt đối với mạng xã hội facebook với nội dung bình luận không hạn chế. Nghiên cứu tích hợp vào mạng xã hội dưới dạng ứng dụng, hỗ trợ cho các tổ chức kinh tế, chính trị, nghệ thuật.

LỜI CÁM TẠ

Nhóm tác giả xin chân thành cảm ơn sự hỗ trợ của khoa CNTT & TT, Trường Đại học Cần Thơ và khoa chuyên ngành trường Cao đẳng Cộng đồng Cà Mau đã tạo điều kiện thuận lợi cho nhóm tác giả hoàn thành đề tài nghiên cứu.

TÀI LIỆU THAM KHẢO

1. Đỗ Thanh Nghị (2011), "Phân loại thư rác với giải thuật ARCX4-rMNB". *Kỷ yếu hội thảo @CNTT – Một số vấn đề chọn lọc của công nghệ thông tin và truyền thông*, pp.427-437.
2. Đỗ Thanh Nghị và Phạm Nguyên Khang (2012), *Nguyên lý máy học*, NXB Đại học Cần Thơ, Cần Thơ.
3. Trần Kim Ngọc (2012), *Phân lớp dữ liệu văn bản lớn dựa trên mô hình túi từ và giải thuật máy học vectơ hỗ trợ SVM*, luận văn thạc sĩ, Trường Đại học Cần Thơ.
4. Adam Bermingham and Alan F. Smeaton. (2010), "Classifying sentiment in microblogs: is brevity an advantage?". In *Proceedings of the 19th ACM international conference on Information and knowledge management* (CIKM '10). ACM, New York, NY, USA, pp. 1833-1836.
5. Bifer, A. and Frank, E. (2010), "Sentiment knowledge discovery in twitter streaming data", *Proceeding of the 13th international conference on Discovery science*, pp. 1-15.
6. Chang, C. and Lin, C-J. (2011). *LIBSVM: A library for support vector machines*. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
7. Go, A., Bhayani, R. and Huang, L. (2009), "Twitter Sentiment Classification using Distant Supervision Technical report".
8. Huifeng Tang, Songbo Tan, and Xueqi Cheng (2009), "A survey on sentiment detection of reviews", *Expert Syst. Appl.* 36, 7 (September 2009), 10760-10773. DOI=10.1016/j.eswa.2009.02.063
9. Lewis, D. and Gale, W (1994), "A sequential algorithm for training test classifiers". In proc, of SIGIR-94.
10. McCallum, A. (1998). *Bow: A Toolkit for Statistical Language Modeling. Text Retrieval, Classification and Clustering*. <http://www2/cs.cmu.edu/~mccallum/bow>
11. Pang, B., Lee, L. and Vaithyanathan, S. (2002). "Thumbs up?: sentiment classification using machine learning techniques", *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, Vol.10, pp. 79-86.
12. Peter Turney (2002), "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews". *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (ACL-02). Jun. 2002, Philadelphia, PN, USA, pp.417–424.