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

KẾT HỢP NGỮ NGHĨA VỚI MÔ HÌNH TÚI TỪ ĐỂ CẢI TIẾN GIẢI THUẬT K LÁNG GIỀNG TRONG PHÂN LỚP VĂN BẢN NGẮN

Đỗ Thanh Nghị[1] và Trần Cao Đệ[1]
[1] *Khoa Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ*

Thông tin chung:
Ngày nhận: 09/05/2014
Ngày chấp nhận: 30/10/2014

Title:
Semantic smoothing of the Bag-of-Words model for improving short text classification using k nearest neighbors

Từ khóa:
Phân lớp văn bản ngắn, mô hình túi từ, ngữ nghĩa, k láng giềng

Keywords:
Text classification, Bag-of-Words, semantic smoothing, k nearest neighbors

ABSTRACT

This paper presents the semantic smoothing of the Bag-of-Words (BoW) model to improve the positive class prediction of k nearest neighbors (kNN) in the short text classification. The BoW model, a representation of the text constructed by counting the occurrence of each word in the text, is popularly used in text classification. The drawback of the BoW model is that it does not take the semantic similarity of words into account. That is often the cause of mismatches in the vocabulary used by kNN. And then, it leads to the poor prediction of the positive class in short text classification. We propose to use the semantic smoothing of BoW to improve the positive class prediction of kNN. The numerical test results on a real dataset show that our approach improves 8% in terms of the positive class prediction while degradesing less than 1% in term of the negative class prediction of kNN algorithm in short text classification.

TÓM TẮT

Trong bài này, chúng tôi giới thiệu tiếp cận tích hợp ngữ nghĩa với mô hình túi từ nhằm cải tiến hiệu quả dự đoán lớp dương của giải thuật k láng giềng trong phân lớp văn bản ngắn. Mô hình túi từ là mô hình biểu diễn văn bản như véc tơ tần số xuất hiện của từ trong văn bản, được sử dụng phổ biến hiện nay trong vấn đề phân lớp văn bản. Tuy nhiên, khuyết điểm của mô hình túi từ là không quan tâm đến sự đồng nghĩa của từ, điều này làm giảm hiệu quả dự đoán lớp dương (lớp quan tâm) của giải thuật k láng giềng trong phân lớp văn bản ngắn. Chúng tôi đề xuất tích hợp ngữ nghĩa vào mô hình túi từ để cải thiện kết quả dự đoán lớp dương của k láng giềng. Kết quả thực nghiệm với tập dữ liệu thực cho thấy rằng các phương pháp của chúng tôi đề xuất cải thiện dự đoán lớp dương hơn 8% trong giảm chưa đến 1% dự đoán lớp âm của giải thuật k láng giềng trong phân lớp văn bản ngắn.

1 GIỚI THIỆU

Phân lớp văn bản (Manning, 2008), (Sebastiani, 99) là gán nhãn tự động cho từng văn bản theo chủ đề đã được định nghĩa trước dựa vào nội dung của văn bản. Phân lớp văn bản sử dụng phổ biến trong ứng dụng như: gán nhãn tự động một bản tin, phân lớp ý kiến người dùng trên các mạng xã hội, trả lời tự động thư điện tử, sắp xếp hộp thư điện tử, nhận dạng thư rác, phân tích nội dung để phát hiện nhóm khủng bố.

Trong bài báo này, chúng tôi xét đến vấn đề phân lớp văn bản ngắn, thường thấy ở các ứng dụng như phân lớp ý kiến trên mạng xã hội twitter (Liu, 2012), kiểm tra các câu hỏi / trả lời từ các

cuộc phỏng vấn trong ứng dụng hoạch định nguồn nhân lực doanh nghiệp (Do *et al*., 2014). Các văn bản này thường rất ngắn (chứa tối đa khoảng 20 từ), mang rất ít thông tin để cho phép thực hiện việc phân lớp bởi các mô hình máy học. Hơn nữa, mô hình túi từ lại không quan tâm đến sự đồng ngữ nghĩa của các từ, tìm hai văn bản tương tự nhau trong giải thuật phân lớp kNN cần phải so khớp từ vựng. Điều này làm giảm hiệu quả dự đoán lớp dương (lớp quan tâm). Chúng tôi đề xuất tích hợp ngữ nghĩa vào mô hình túi từ để cải thiện kết quả dự đoán lớp dương của kNN trong phân lớp văn bản ngắn. Mô hình ngữ nghĩa dựa trên tự điển đồng nghĩa WordNet (Fellbaum, 1998), phân tích ngữ nghĩa tiềm ẩn LSA (Dumais, 2004) và chủ đề tiềm ẩn LDA (Blei *et al*., 2003). Kết quả thực nghiệm với tập dữ liệu thu được từ nghiên cứu (Do *et al*., 2014) cho thấy rằng các phương pháp của chúng tôi đề xuất cải thiện dự đoán lớp dương hơn 8% trong khi giảm chưa đến 1% dự đoán lớp âm của giải thuật kNN trong phân lớp văn bản ngắn.

Phần tiếp theo của bài viết được trình bày như sau: phần 2 trình bày ngắn gọn về phân lớp văn bản với mô hình túi từ và giải thuật kNN, phần 3 trình bày các phương pháp kết hợp ngữ nghĩa với mô hình túi để cải tiến giải thuật phân lớp kNN. Phần 4 trình bày các kết quả thực nghiệm, tiếp theo sau đó là thảo luận về các nghiên cứu có liên quan đến phân lớp văn bản trước khi kết luận và hướng phát triển.

2 PHÂN LỚP VĂN BẢN VỚI MÔ HÌNH TÚI TỪ VÀ GIẢI THUẬT KNN

Phương pháp phân lớp văn bản thường dựa trên mô hình thống kê từ và các giải thuật học tự động (Manning, 2008), (Sebastiani, 99).

2.1 Mô hình túi từ

Do dữ liệu văn bản ở đầu vào ở dạng không cấu trúc, trong khi các giải thuật máy học ở giai đoạn tiếp theo sau thường chỉ có thể xử lý được dữ liệu dạng cấu trúc bảng (mỗi dòng là một phần tử dữ liệu, cột là chiều hay thuộc tính). Để giải quyết vấn đề này, mô hình túi từ (Harris, 1954), (Salton *et al*., 1975) cho phép chúng ta biểu diễn tập dữ liệu văn bản về cấu trúc bảng.

Bảng 1: Ví dụ về tập dữ liệu văn bản

STT	Nội dung	Lớp
1	support vector machines for classifying images	Dương
2	databases management systems	Âm
…	…	…
m	software testing	Âm

Bước tiền xử lý này bao gồm việc phân tích từ vựng và tách các từ trong nội dung của tập văn bản, sau đó chọn tập hợp các từ có ý nghĩa quan trọng dùng để phân lớp, biểu diễn dữ liệu văn bản về dạng bảng để từ đó các giải thuật máy học có thể học để phân lớp. Ở bước phân tích từ vựng, công việc có thể là quy về từ gốc của các biến thể từ, có thể xóa bỏ các từ không có ý nghĩa cho việc phân lớp như các mạo từ, từ nối,... Tiếp đến là tách các từ, đưa vào tự điển. Một văn bản được biểu diễn dạng véc tơ (có n thành phần, chiều) mà giá trị thành phần thứ j là tần số xuất hiện từ thứ j trong văn bản. Nếu xét tập T gồm m văn bản và tự điển có n từ vựng, thì T có thể được biểu diễn thành bảng D kích thước m×n, dòng thứ i của bảng là véc tơ biểu diễn văn bản thứ i tương ứng.

Xem ví dụ tập dữ liệu văn bản trong Bảng 1, sau khi tiền xử lý biểu diễn với mô hình túi từ thu được bảng dữ liệu D có cấu trúc như Bảng 2, từ bảng này giải thuật máy học như kNN có thể xử lý vấn đề phân lớp.

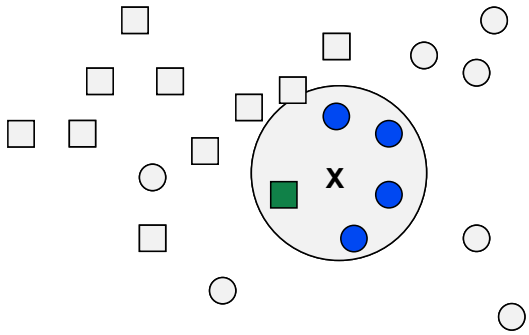
Bảng 2: Biểu diễn tập dữ liệu văn bản bằng mô hình túi từ

STT	1 (support)	2 (machine)	…	n-1 (software)	n (system)	Lớp
1	1	1	…	0	0	Dương
2	0	0	…	0	1	Âm
…	…	…	…		…	…
m	0	0	…	1	0	Âm

2.2 Giải thuật kNN

Giải thuật kNN được Fix và Hodges đề xuất từ những năm 1952. Đây là phương pháp rất đơn giản nhưng cũng cho hiệu quả cao trong khai mỏ dữ liệu (Wu and Kumar, 2009). Giải thuật kNN thường được sử dụng trong các vấn đề tìm kiếm văn bản, phân lớp văn bản (Manning, 2008). Giả sử ta có tập mẫu dữ liệu học ban đầu có 2 lớp (tròn, vuông) như ví dụ trong Hình 1. Giải thuật kNN không cần quá trình học. Khi có một phần tử dữ liệu *x* mới đến cần dự đoán lớp, giải thuật đi tìm trong tập học *k*

láng giềng từ tập dữ liệu học ($k = 5$) của phần tử mới đến x để thực hiện dự đoán, lớp của phần tử mới đến x được dự đoán dựa vào luật bình chọn số đông từ các lớp của k láng giềng (lớp của phần tử x được dự đoán là tròn).



Hình 1: Giải thuật kNN

Bảng 3: Ví dụ về tập dữ liệu 2 văn bản *d1*, *d2* và văn bản *x* cần phân lớp

STT	Nội dung	Lớp
d1	kernel machines	Dương
d2	Linux machine	Âm
x	support vector machines	?

Trong khi phân lớp, giải thuật kNN cần tìm trong tập học k láng giềng. Chính vì vậy, kết quả phân lớp của giải thuật phụ thuộc vào độ đo khoảng cách. Tuy nhiên, nếu sử dụng mô hình túi từ như trên, thì khi tìm k láng giềng của kNN lại không thể xét đến sự tương đồng về mặt ngữ nghĩa do mô hình túi từ không quan tâm đến sự đồng nghĩa của các từ trong văn bản (mà cần so khớp chính xác từ). Để thấy rõ điều này, cần xem xét ví dụ với tập dữ liệu có 3 văn bản ngắn (2 văn bản d1, d2 được gán sẵn lớp và văn bản x cần phân lớp) như sau Bảng 3.

Bảng 4: Mô hình túi từ của tập dữ liệu 2 văn bản *d1*, *d2* và văn bản *x* cần phân lớp

STT	support	vector	machine	linux	kernel	Lớp
d1	0	0	1	0	1	Dương
d2	0	0	1	1	0	Âm
x	1	1	1	0	0	?

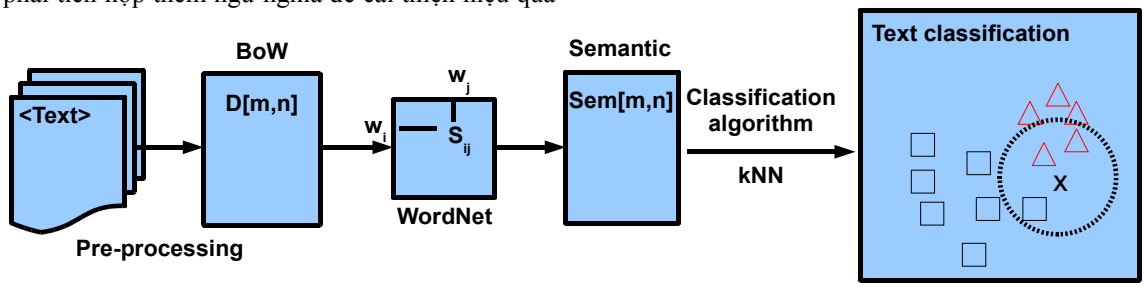
Bảng 4 là mô hình túi từ biểu diễn 3 văn bảng trên. Để phân lớp x thuộc vào lớp *Dương* hay *Âm*, kNN tìm *1* láng giềng của x bằng cách tính khoảng cách (chẳng hạn Manhattan) từ x đến văn bản *d1* và *d2*, sau đó láng giềng là khoảng cách nhỏ nhất. Kết quả là khoảng cách $d(x, d1) = d(x, d2) = 3$. kNN không thể phân lớp tốt cho phần tử x. Nếu xét về ngữ nghĩa, thì x tương tự với *d1* hơn là *d2*, do x và *d1* cùng thuộc vào chủ đề máy học.

3 TÍCH HỢP NGỮ NGHĨA VÀO MÔ HÌNH TÚI TỪ

Để khắc phục nhược điểm về việc bỏ qua sự đồng nghĩa của từ trong mô hình túi từ, cần thiết phải tích hợp thêm ngữ nghĩa để cải thiện hiệu quả phân lớp của giải thuật kNN. Có hai tiếp cận được đề xuất ở đây. Một là sử dụng tự điển đồng nghĩa WordNet (Fellbaum, 1998) kết hợp với mô hình túi từ. Ngoài ra, có thể sử dụng phương pháp phân tích ngữ nghĩa tiềm ẩn (Latent Semantic Analysis – LSA (Dumais, 2004) và chủ đề tiềm ẩn Latent Dirichlet Allocation - LDA (Blei *et al.*, 2003)).

3.1 Mô hình ngữ nghĩa sử dụng tự điển WordNet

Xuất phát từ mô hình túi từ với tập từ vựng Voc kích thước là n, cần xây dựng một ma trận đồng nghĩa S kích thước là n×n mà ở đó mỗi phần tử Sij là mối quan hệ đồng nghĩa giữa từ Wi và từ Wj trong tập Voc.



Hình 2: Mô hình ngữ nghĩa BoW-WordNet và kNN cho phân lớp văn bản

Ma trận đồng nghĩa S được xây dựng dựa vào tự điển WordNet. Tự điển này là một cơ sở dữ liệu ngữ nghĩa phân cấp của các từ tiếng Anh, cung cấp các mối quan hệ đồng nghĩa giữa các từ. Sự đồng

nghĩa ngữ nghĩa giữa hai từ có giá trị trong đoạn *[0, 1]*. Nếu hai từ là đồng nghĩa tuyệt đối thì WordNet trả về giá trị đồng nghĩa ngữ nghĩa là *1*.

Việc tích hợp đồng nghĩa vào mô hình túi từ thực hiện bằng cách nhân bảng dữ liệu *D* kích thước $m \times n$ của mô hình túi từ với ma trận đồng nghĩa *S*, thu được bảng dữ liệu mới là *Sem* có kích thước là $m \times n$. Lúc này giải thuật kNN xử lý bảng *Sem* để phân lớp dữ liệu thay vì là bảng *D* của mô hình túi từ thường thấy. Hình 2 mô tả mô hình ngữ nghĩa BoW-WordNet và kNN cho phân lớp văn bản.

Trở lại ví dụ trên, với bộ từ vựng gồm 5 từ {support; vector; machine; linux; kernel}, ma trận đồng nghĩa *S* được xây dựng từ tự điển WordNet như trình bày trong Bảng 5.

Bảng 5: Ma trận đồng nghĩa *S* của tập từ vựng {support; vector; machine; linux; kernel}

S	support	Vector	machine	linux	Kernel
support	1.0	0.38	0.78	0.30	0.39
vector	0.38	1.0	0.38	0.21	0.47
machine	0.78	0.38	1.0	0.0	0.36
linux	0.30	0.21	0.0	1.0	0.0
kernel	0.39	0.47	0.36	0.0	1.0

Tiếp đến, nhân bảng dữ liệu 4 của mô hình túi từ với ma trận đồng nghĩa *S* ở Bảng 5, thu được bảng dữ liệu *Sem* như Bảng 6. Nếu để ý so sánh sự khác nhau giữa Bảng 4 (mô hình túi từ) và Bảng 6 (mô hình túi từ có tích hợp ngữ nghĩa), ta có thể thấy rằng trong Bảng 4 véc tơ *x* có thành phần tương ứng *kernel = 0* do *x* không có chứa từ *kernel*, tuy nhiên các từ *support*, *vector*, *machine* lại có quan hệ ngữ nghĩa từ *kernel* nên sau khi tích hợp ngữ nghĩa, véc tơ *x* trong với Bảng 6 có thành phần *kernel* tương ứng là giá trị lớn hơn *0*. Tương tự như vậy, nhiều giá trị *0* của Bảng 4 cũng được thay bằng các giá trị lớn hơn *0* trong Bảng 6 do từ tương ứng có quan hệ ngữ nghĩa với các từ khác xuất hiện trong văn bản.

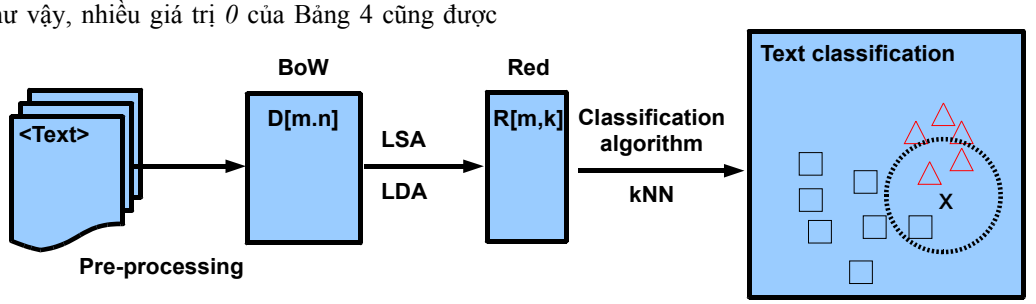
Bảng 6: Mô hình túi từ đã tích hợp ngữ nghĩa *Sem* của tập dữ liệu {*d1*, *d2* và *x*}

STT	support	vector	machine	Linux	kernel	Lớp
d1	1.17	0.85	1.36	0	1.36	Dương
d2	1.08	0.58	1	1	0.36	Âm
x	2.16	1.76	2.16	0.51	1.21	?

Để phân lớp văn bản *x* sử dụng bảng 6 (bảng dữ liệu đã tích hợp ngữ nghĩa *Sem*), kNN tìm 1 láng giềng của *x* bằng cách tính khoảng cách Manhattan từ *x* đến văn bản *d1* và *d2*. Kết quả là khoảng cách $d(x, d1) = 3.35 < d(x, d2) = 4.76$. Văn bản *x* được gán lớp là *Dương* (lớp của *d1*). Kết quả này chứng tỏ rằng tích hợp ngữ nghĩa vào mô hình túi từ giúp kNN phân lớp hiệu quả văn bản hơn.

3.2 Phân tích ngữ nghĩa tiềm ẩn (LSA)

Phân tích ngữ nghĩa tiềm ẩn (LSA) là một kỹ thuật thường dùng trong xử lý ngôn ngữ tự nhiên và tìm kiếm thông tin. LSA thực hiện phân tích các mối quan hệ giữa tập các văn bản và các từ vựng có trong văn bản. LSA giả định rằng những từ có ngữ nghĩa gần nhau thường xuất hiện trong cùng ngữ cảnh (phần tương tự của văn bản). Xuất phát từ bảng dữ liệu *D* kích thước $m \times n$ (*m* văn bản, *n* từ vựng) thu được ở mô hình túi từ, LSA sử dụng kỹ thuật phân tích giá trị kỳ dị (Singular Value Decomposition - SVD), rút trích mối tương quan ngữ nghĩa giữa các từ trong tập văn bản, giảm số cột (chiều) về *k* đặc trưng tiềm ẩn của bảng dữ liệu, thu được bảng *R* kích thước $m \times k$ trong khi vẫn giữ được cấu trúc tương tự của các dòng trong bảng *R*. Giải thuật kNN sử dụng bảng *R* (ngữ nghĩa tiềm ẩn) để thực hiện phân lớp văn bản. Mô hình được mô tả như Hình 3.



Hình 3: Mô hình ngữ nghĩa tiềm ẩn BoW-LSA/LDA và kNN cho phân lớp văn bản

3.3 Chủ đề ẩn (LDA)

Trong khi LSA chỉ có thể rút trích ngữ nghĩa đồng nghĩa tiềm ẩn của các từ trong tập văn bản thì một vấn đề khác vẫn chưa được giải quyết đó chính là sự đa nghĩa của từ. Để giải quyết được cả vấn đề trên, (Blei *et al.*, 2003) đã đề xuất mô hình chủ đề

tiềm ẩn LDA cho phép khám phá ngữ nghĩa tiềm ẩn và cả sự đa nghĩa của từ trong tập văn bản. LDA là một mô hình sinh xác suất cho tập dữ liệu rời rạc (văn bản). LDA giả định rằng mỗi tài liệu là sự trộn lẫn của nhiều chủ đề, mỗi chủ đề là một phân bố xác suất trên các từ. LDA về bản chất được xem là mô hình Bayes 3 cấp độ: tập văn bản, văn bản và từ; trong đó mỗi văn bản của tập hợp *m* văn bản được mô hình như một mô hình hỗn hợp của *k* chủ đề ẩn, mỗi chủ đề ẩn là phân phối xác suất đa thức (multinomial distribution) của *n* từ. Mô hình LDA sinh từ và văn bản theo quy tắc như sau:

Ứng với mỗi từ của n_j từ vựng của văn bản d_j,

– lấy mẫu chủ đề z_{ij} tuân theo phân phối xác suất đa thức *multinomial*(θ_j)

– lấy mẫu từ w_{ij} tuân theo phân phối xác suất đa thức *multinomial*(ϕz_{ij})

Ở đó những tham số của phân phối xác suất đa thức cho chủ đề trong một tài liệu là θ_j và những từ trong chủ đề ẩn ϕk có xác suất tiên nghiệm Dirichlet. Một cách trực quan, có thể diễn giải về tham số ϕk để chỉ tầm quan trọng của những từ trong chủ đề k và tham số θ_j chỉ ra những chủ đề khám phá trong văn bản d_j.

Xuất phát từ bảng dữ liệu *D* kích thước $m \times n$ (*m* văn bản, *n* từ vựng) thu được ở mô hình túi từ, LDA rút trích *k* chủ đề tiềm ẩn trong tập *m* văn bản, tạo bảng dữ liệu mới *R* kích thước $m \times k$. Giải thuật kNN sử dụng bảng *R* (chủ đề tiềm ẩn) để thực hiện phân lớp văn bản (xem Hình 3).

4 KẾT QUẢ THỰC NGHIỆM

Phần thực nghiệm nhằm đánh giá hiệu quả của tiếp cận đề xuất sử dụng ngữ nghĩa kết hợp với mô hình túi từ để cải tiến hiệu quả dự đoán lớp dương của giải thuật kNN trong phân lớp văn bản ngắn.

Về chương trình, chúng tôi đã tiến hành cài đặt giải thuật kNN, phương pháp phân tích ngữ nghĩa tiềm ẩn LSA và chủ đề ẩn LDA bằng ngôn ngữ lập trình C/C++ trên máy PC Linux.

Bảng 7: Ví dụ về tập dữ liệu văn bản gồm các câu hỏi / trả lời từ các cuộc phỏng vấn trong ứng dụng hoạch định nguồn nhân lực doanh nghiệp

STT	Nội dung	Lớp
1	Q: Are all your employees working at the same building? A: Inconsistent C:	Dương
2	Q: What kind of clients does the company have? A: C: Check for update	Âm
…	…	…

Chúng tôi sử dụng tập dữ liệu văn bản ngắn được nghiên cứu trong (Do *et al*., 2014). Đây là tập dữ liệu văn bản gồm các câu hỏi / trả lời từ các cuộc phỏng vấn trong ứng dụng hoạch định nguồn nhân lực doanh nghiệp. Các văn bản ngắn này thường rất ngắn (chứa khoảng 20 từ), như ví dụ trong Bảng 7. Vấn đề cần xử lý ở đây là dựa vào nội dung câu hỏi (Q) và câu trả lời (A) mà kiểm tra xem cặp Q/A được phân lớp là đúng (dương) hay sai (âm) cần chỉnh sửa (C). Tập dữ liệu đầu tiên có 3696 cặp Q/A, sau bước tiền xử lý loại bỏ các cặp Q/A trùng lặp, chúng tôi chỉ còn lại 966 cặp Q/A. Chúng tôi sử dụng thư viện Libbow của (McCallum, 1998) để thực hiện bước tiền xử lý dữ liệu văn bản (loại bỏ từ ít ý nghĩa stop-words và quy về từ gốc), xây dựng mô hình túi từ với 500 từ. Chúng tôi thu được tập dữ liệu là bảng *D* có 966 dòng (cặp Q/A) với 500 cột (chiều, từ), có 2 lớp (dương/âm) trong đó 112 phần tử thuộc lớp dương (chiếm tỷ lệ khoảng 12%) và 854 phần tử thuộc lớp âm (chiếm tỷ lệ khoảng 88%).

Nghi thức kiểm tra được sử dụng ở đây là hold-out, lặp lại 10 lần, mỗi lần lấy ngẫu nhiên 644 dòng làm tập học và 322 dòng còn lại làm tập kiểm tra. Tính hiệu quả là trung bình của 10 lần lặp. Giải thuật kNN đạt hiệu quả cao nhất với *k=3*.

kNN phân lớp bảng dữ liệu thu được từ mô hình túi từ được ký hiệu là kNN.

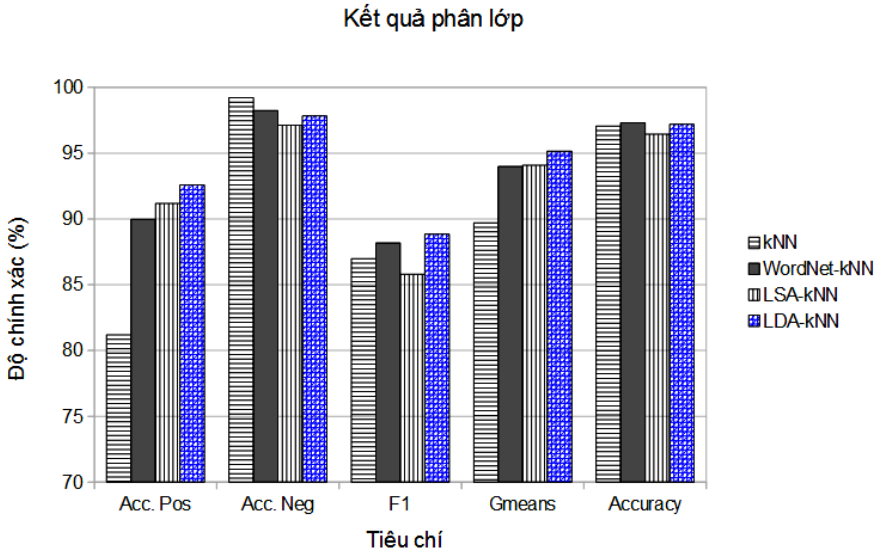
Riêng phần tích hợp ngữ nghĩa WordNet với mô hình túi từ, chúng tôi sử dụng thư viện cung cấp bởi (Seco *et al*., 2004), kNN phân lớp trên bảng dữ liệu này được ký hiệu là WordNet-kNN.

Chúng tôi dùng LSA và LDA để rút trích tương ứng 50 đặc trưng tiềm ẩn và sử dụng kNN trên các bảng dữ liệu này được ký hiệu lần lượt là LSA-kNN và LDA-kNN.

Chúng tôi tiến hành so sánh kết quả dựa trên các tiêu chí như độ chính xác của lớp dương (Acc. Pos), độ chính xác của lớp âm (Acc. Neg), Gmeans (trung bình điều hòa giữa độ chính xác của lớp

dương, lớp âm), độ chính xác toàn cục (Accuracy), độ đo F1 (van Rijsbergen, 1979). Accuracy là số phần tử được phân lớp đúng của tất cả các lớp chia cho tổng số phần tử. F1 là trung bình điều hòa của precision và recall; trong đó precision là số phần tử được phân lớp đúng về lớp dương chia cho tổng số phần tử được dự đoán là lớp dương và recall là số phần tử được phân lớp đúng về lớp dương này chia cho tổng số phần tử của dương.

Kết quả thu được từ các giải thuật được trình bày trong Hình 4. Quan sát kết quả thu được Acc. Pos, có thể thấy rằng tất cả các mô hình sử dụng ngữ nghĩa WordNet, LSA, LDA đều cải thiện kết quả dự đoán lớp dương trên 8% so với sử dụng kNN trực tiếp trên mô hình túi từ. Tuy nhiên, chỉ có WordNet và LDA vẫn còn duy trì kết quả dự đoán lớp âm trong khi LSA làm giảm dự đoán lớp âm đến 2% so với mô hình túi từ gốc. Điều này làm cho LSA-kNN cho kết quả thấp khi so tiêu chí F1 và Accuracy. Kết quả của LSA-kNN xuất phát từ nguyên nhân LSA chưa giải quyết tốt cho vấn đề đa nghĩa của từ. Trong khi đó 2 mô hình ngữ nghĩa còn lại WordNet-kNN và LDA-kNN lại luôn chiếm ưu thế khi so sánh trên F1, Gmeans và Accuracy.



Hình 4: Kết quả phân lớp tập văn bản ngắn

Kết quả thu được từ thực nghiệm này cho phép chúng tôi tin rằng tích hợp ngữ nghĩa vào mô hình túi từ WordNet-kNN và LDA-kNN cải thiện được đáng kể hiệu quả dự đoán lớp dương và giảm rất ít kết quả dự đoán lớp âm của giải thuật kNN trong phân lớp văn bản ngắn.

5 THẢO LUẬN VỀ CÁC NGHIÊN CỨU LIÊN QUAN

Các tiếp cận phân lớp văn bản được nghiên cứu trước đây dựa trên mô hình ngữ nghĩa hoặc máy học (Manning, 2008), (Sebastiani, 1999). Theo giáo sư đầu ngành về phân tích dữ liệu của Đại học California, Berkeley, M. Hearst cho rằng tiếp cận ngữ nghĩa rất phức tạp và không mang nhiều hiệu quả. Thay vì vậy, tiếp cận dựa trên máy học tự động lại đơn giản và cho nhiều kết quả tốt trong thực tiễn. Phương pháp phân lớp văn bản dựa trên mô hình thống kê từ và các giải thuật học tự động. Dữ liệu văn bản có độ dài khác nhau được biểu diễn dưới dạng véc tơ tần số xuất hiện của từ trong văn bản (mô hình túi từ (Harris, 1954), (Salton *et al*., 1975)), đây là mô hình biểu diễn phổ biến và được dùng trong hầu hết các nghiên cứu về phân lớp văn bản và tìm kiếm thông tin, (Manning, 2008), (Sebastiani, 1999), (Lewis & Gale, 1994). Tập từ vựng thu được có thể lên đến hàng trăm ngàn. Vì vậy, tập dữ liệu văn bản được chuyển về dạng một bảng có số cột (chiều, từ vựng) rất lớn. Bước tiếp theo là huấn luyện mô hình học tự động từ bảng dữ liệu này. Các mô hình máy học thường sử dụng như giải thuật k láng giềng (kNN (Fix & Hodges, 52)), Bayes thơ ngây (NB (Good, 65)), cây quyết định (Quinlan, 93), máy học véc tơ hỗ trợ (SVM (Vapnik, 95)), giải thuật tập hợp mô hình bao gồm Boosting (Freund & Schapire, 95) và rừng ngẫu nhiên (Breiman, 01).

Nghiên cứu giảm chiều với phân tích ngữ nghĩa tiềm ẩn và chủ đề ẩn cũng được tìm thấy trong các công trình tiêu biểu như (Hofmann, 1999),

(Dumais, 2004), (Blei *et al*., 2003), (Trần & Phạm, 2012).

Cũng đã có những nghiên cứu trước đây của chúng tôi trong (Phạm *et al*., 2006, 2008), (Đỗ & Phạm, 2013), đề xuất giải thuật tập hợp mô hình của máy học SVM, Bayes thơ ngây, cây xiên ngẫu nhiên, cho phân lớp hiệu quả dữ liệu văn bản biểu diễn trực tiếp từ mô hình túi từ có số chiều lớn. Ngoài ra, chúng tôi cũng đã đề xuất tích hợp ngữ nghĩa bằng tự điển WordNet (Fellbaum, 1998), cho phép cải thiện kết quả hiển thị và tìm kiếm chuyên gia (Nguyen *et al*., 2009) và tìm kiếm văn bản (Bùi *et al*., 2006).

Tuy nhiên, nghiên cứu của bài viết được đặt trong ngữ cảnh phân lớp văn bản ngắn (chứa rất ít từ) sử dụng mô hình túi từ và máy học dựa trên khoảng cách đơn giản như kNN. (Song *et al*., 2014) nêu bật vấn đề khó khăn khi xử lý văn bản ngắn và các tiếp cận máy học như chủ đề ẩn, máy học SVM, kNN, NB. Mục tiêu không nhằm so sánh với các giải thuật máy học phức tạp khác mà ý tưởng chính là cải tiến tiếp cận thường thấy trong phân lớp văn bản và tìm kiếm thông tin có sử dụng mô hình túi từ biểu diễn văn bản và giải thuật dựa trên cách tính khoảng cách luôn bỏ qua sự đồng nghĩa của từ.

6 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Chúng tôi vừa trình bày tiếp cận tích hợp ngữ nghĩa nhằm nâng cao hiệu quả phân lớp văn bản ngắn của giải thuật kNN. Văn bản ngắn mang rất ít thông tin để cho phép thực hiện việc phân lớp bởi kNN. Trong khi mô hình túi từ để biểu diễn văn bản hiện nay lại không quan tâm đến sự đồng ngữ nghĩa của các từ, hai văn bản tương tự nhau trong giải thuật kNN cần phải so khớp từ vựng. Điều này làm giảm hiệu quả dự đoán lớp dương (lớp quan tâm). Để cải thiện kết quả dự đoán lớp dương của kNN trong phân lớp văn bản ngắn, chúng tôi đề xuất tích hợp ngữ nghĩa vào mô hình túi từ, sử dụng tự điển đồng nghĩa WordNet, ngữ nghĩa tiềm ẩn LSA và chủ đề ẩn LDA. Kết quả thực nghiệm với tập dữ liệu thực tiễn văn bản ngắn cho thấy rằng các phương pháp của chúng tôi đề xuất cải thiện dự đoán lớp dương hơn 8% trong khi giảm chưa đến 1% dự đoán lớp âm của giải thuật kNN trong phân lớp văn bản ngắn.

Trong tương lai, chúng tôi dự định mở rộng ý tưởng này để xử lý vấn đề tương tự như tìm kiếm, phân lớp, ảnh, videó, có sử dụng mô hình biểu diễn túi từ và mô hình máy học dựa trên khoảng cách như kNN.

TÀI LIỆU THAM KHẢO

1. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. *Journal of Machine Learning Research* 3 (4–5): 993-1022, (2003).
2. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001).
3. Bùi T-T., Nguyễn Đ-T. và Đỗ T-N.: Hệ thống tư vấn tài nguyên học tập. Kỷ yếu hội thảo SGK'06, Huế, Tr. 1-9 (2006).
4. Do, T-N., Moga, S. and Lenca, P.: Random forest of oblique decision trees for ERP semi-automatic configuration. in Multiple Model Approach to Machine Learning, Springer (2014), pp. 25-34.
5. Đỗ, T-N., Phạm, N-K.: Phân loại văn bản: Mô hình túi từ và tập hợp mô hình máy học tự động. *Tạp chí khoa học ĐHCT*, Số 28: 9-16 (2013).
6. Dumais, S.: Latent Semantic Analysis. *Annual Review of Information Science and Technology* Vol. 38(1):188-230, (2004).
7. Fellbaum, C.: *WordNet: An electronic lexical database*. MIT Press (1998)
8. Fix, E. and Hodges J.: Discriminatoiry Analysis: Small Sample Performance. Technical Report 21-49-004, USAF School of Aviation Medicine, Randolph Field, USA (1952).
9. Freund, Y., and Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. In: Computational Learning Theory: Proceedings of the Second European Conference, pp. 23–37 (1995).
10. Good, I.: The Estimation of Probabilities: An Essay on Modern Bayesian Methods. *MIT Press* (1965).
11. Harris, Z.: Distributional Structure. *Word* 10(2/3) (1954).
12. Hofmann, T.: Probabilistic Latent Semantic Indexing. Proceedings of the 22th Annual International SIGIR Conference on Research and Development in Information Retrieval (1999), pp.
13. Lewis, D., Gale, W.: A sequential algorithm for training text classifiers. In: Proceedings of SIGIR (1994).
14. Liu, B.: *Sentiment Analysis and Opinion Mining*. Morgan & Claypool, 2012.

15. Manning, C., Raghavan, P. and Schütze, H.: *Introduction to Information Retrieval*. Cambridge University Press (2008)
16. McCallum, A.: Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering. 1998. http://www-2.cs.cmu.edu/~mccallum/bow.
17. Nguyen, T-B., Lenca, P., Do, T-N. et Poulet, F.: Visualisation de réseaux d'experts. Acte du 7ème Atelier Visualisation et extraction de connaissances, EGC'09, 9èmes Journées d'Extraction et Gestion des Connaissances (2009), pp. 1-5.
18. Phạm, N-K, Đỗ, T-N, Poulet, F.: Phân loại văn bản với giải thuật Boosting PSVM. Kỷ yếu hội nghị @CNTT (2006), Tr. 269-278.
19. Phạm, N-K., Đỗ, T-N., Trần, C-Đ.: Phân loại dữ liệu với Giải thuật Arcx4-LSSVM. Tuyển tập công trình nghiên cứu Công nghệ Thông tin và Truyền thông, NXB KHKT (2008), Tr.72-78.
20. Quinlan, J.R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA (1993).
21. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. *Communications of the ACM*, Vol.18(11):613-620 (1975).
22. Sebastiani, F.: Machine learning in automated text categorization. *ACM Computing Surveys* 34(1), 1–47 (1999)
23. Seco, N., Veale, T., Hayes, J.: An Intrinsic Information Content Metric for Semantic Similarity in WordNet. Proceedings of ECAI (2004), pp. 1089-1090.
24. Song, G., Ye, Y., Du, X., Huang, X., Bie, S.: Short Text Classification: A Survey. *Journal of Multimedia*, Vol.9(5):635-643 (2014).
25. Trần, C.Đ và Phạm N.K.: Phân loại văn bản với máy học véc tơ hỗ trợ và cây quyết định. *Tạp chí khoa học ĐH. Cần Thơ* số (21a):52-63 (2012).
26. Van Rijsbergen, C.V.: *Information Retrieval*. Butterworth (1979)
27. Vapnik, V.: *The Nature of Statistical Learning Theory*. Springer-Verlag (1995)
28. Wu, X. and Kumar, V.: *Top 10 Algorithms in Data Mining*. Chapman & Hall/CRC (2009).