TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

Phân tích quan điểm cho dữ liệu mạng xã hội tiếng Việt

Lê Thanh Hương
Viện Công nghệ Thông tin và Truyền thông, ĐHBKHN

Mở đầu

- MXH xuất hiện đầu những năm 2000 nhằm chia sẻ quan điểm của họ về sự kiện, con người, sản phẩm
  - Nguồn thông tin quí giá trong kinh doanh
- Phân tích quan điểm với văn bản dài → SVM
- PTQĐ với câu ngắn từ MXH → khó hơn

PTQĐ trên MXH

- Vd 1: *Ha-ha... I want to see. E macdonalds here cheaper. Yum.* → *positive*

- Vd 2: *Ya... She wans... But now so late dunno still can arrange 4 tmr anot...* → *negative*
  - Vd. 2 chứa nhiều từ ko chính thống: "ya", "wans", "dunno", "4", "tmr", "anot"

PTQĐ trên MXH

Nhạc trẻ Hot 5 tháng trước

Hay Tuyệt chị mỹ tâm ^^ --> POSITIVE

34 TRẢ LỜI

Nguyễn Hương 1 tháng trước

Giống câu chuyện người con gái nuôi người yêu đến khi anh ta thành đạt thì bỏ chị ấy. --> NEUTRAL

4 TRẢ LỜI

Thái Bình Nguyễn 3 năm trước

Dáng đi xấu, hát ko ra hồn bài hát. Nghiêm túc mà nói quá tệ. --> NEGATIVE

9 TRẢ LỜI

Các khó khăn trong PTQĐ MXH Việt

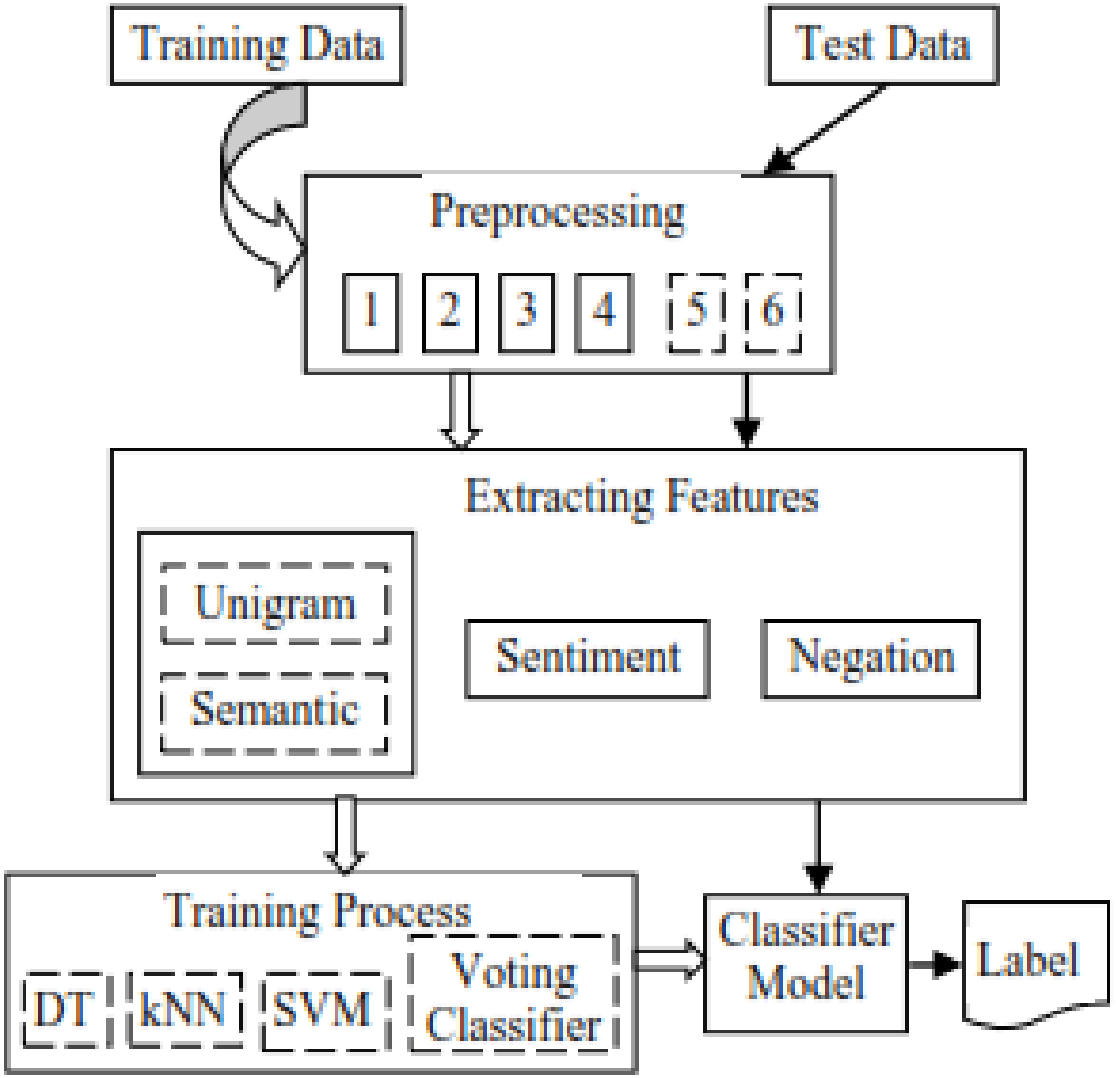
- Ngôn ngữ không chính thống, nhiều biểu tượng, từ lóng, lỗi ngữ pháp
- Chưa có bộ dữ liệu tiếng Việt để huấn luyện
- Chưa có nhiều nghiên cứu về PTQĐ MXH tiếng Việt → khó so sánh kết quả

=> Nghiên cứu và cài đặt cho bộ dữ liệu tiếng Anh trước khi làm cho tiếng Việt

Các nghiên cứu liên quan

- Kiritchenko et al [3]: linear-kernel SVM với đặc trưng ngram, ngữ nghĩa, từ quan điểm, từ phủ định
- Deshwal and Sharma [2]: 6 thuật toán học có giám sát với đặc trưng biểu tượng, dấu câu (?!), từ điển từ, unigrams
- Lango et al.[8]: Random Forests, SVMs, Gradient Boosting Trees, đặc trưng ngrams, phân cụm Brown, từ điển cảm xúc, WorldNet, POS
- Aueb[6]: học giám sát kết hợp có trọng số các bộ phân loại, sử dụng GloVe word embeddings cho dữ liệu Twitter

Kiến trúc hệ thống



Biểu diễn dữ liệu

- Word embedding: FastText (huấn luyện với Wikipedia)
- Huấn luyện lại với bộ từ Sentiment140 (dữ liệu Twitter), 300 chiều
- từ Sentiment140 có nhiều từ mở rộng như "hello“ → "hellllooooo"
- Tập dữ liệu được tiền xử lý bằng cách thay >= 3 ký tự liên tiếp = 2 (vd, niccccceeeee → niccee) trước khi huấn luyện. Mục đích: giảm kích thước bộ từ vựng

Tiền xử lý

1. Chuyển xâu về chữ thường
2. Chuyển link → URL, @username → AT_USER;
3. Chuyển chữ viết tắt, tiếng lóng, emoticons thành ý nghĩa của nó: :) → happy, “dunno” → “don’t know”
4. Loại bỏ dấu cách thừa
5. Thay >= 3 ký tự liên tiếp = 2 (vd, niccccceeeee → niccee)
6. Trích mệnh đề chính trong câu có quan hệ đối lập

Tiền xử lý

- Chuyển chữ viết tắt, từ lóng, emoticons thành ý nghĩa của nó:
  - Sử dụng từ điển Twitter
  - Sử dụng word embedding để lấy ý nghĩa của chữ viết tắt, từ lóng
- Thay >= 3 ký tự liên tiếp = 2 (vd, niccccceeee → niccee)
  - Cách khác: phân cụm (yes, yeesss, yep), word embedding
- Trích mệnh đề chính trong câu có quan hệ đối lập
  - VD: "*I thought it was good, but it was awful.*" → *negative*
  - Sử dụng từ hiệu để phân biệt

Đặc trưng

1. Word unigrams: bộ từ vựng lớn do tiếng lóng,…
2. Ngữ nghĩa: FastText word embedding
3. Từ cảm xúc: SentiWordNet, mỗi từ nhận giá trị cảm xúc positivity, negativity, $\in$ 0÷1

- Phủ định: not, cant, never

Bộ phân loại

- kNN (k=24), SVM, cây quyết định, Voting classifier trong scikit-learn

$$y_{\text{Voting Classifier}} = \text{argmax}_v(\sum i\ w_i * p_{i,v})$$

Wi là trọng số của bộ phân loại

$$\sum_v p_{i,v} = 1 \quad \text{với } \forall \text{ i}$$

Dữ liệu thử nghiệm

- 3 bộ Twitters datasets: từ Sentiment140, Twitter 2013 trong SemEval2013, Twitter 2016 trong SemEval2016 cho task 4, subtask A
- Train word2vec: từ Sentiment140 (1.6 triệu tweets)
- Train bộ phân loại quan điểm: Twitter 2013 + Twitter 2016 training và developing dataset: 19337 tweets: 8152 dương, 8133 trung tính, 3052 âm
- Test: 3547 tweets từ Twitter 2013 test

Kết quả thí nghiệm

Bộ phân loại	F1 (%)
Cây quyết định	52.2
kNN	57.0
SVM	59.6
Voting Classifier	637

Kịch bản thử nghiệm

1. ∀ bước tiền xử lý + unigram + từ quan điểm + từ phủ định
2. ∀ bước TXL+ ngữ nghĩa + từ quan điểm + từ phủ định
3. ∀ bước TXL + ngữ nghĩa + từ quan điểm
4. bước TXL1,2,3,4,6 + ngữ nghĩa + từ quan điểm + từ phủ định
5. bước TXL 1,2,3,4,5 + ngữ nghĩa + từ quan điểm + từ phủ định

Kịch bản	1	2	3	4	5
F1 (%)	55.2	63.7	58.3	58.5	63.5

So với các hệ thống khác

	Xếp hạng trong SemEval	F1 (%)
Switchcheese [12]	1	63.3
Unimelb [5]	3	61.7
Aueb [6]	5	60.5
PUT [8]	14	57.6
Our system		63.7

Giải pháp với tiếng Việt

- Xây dựng tập dữ liệu huấn luyện:
  - Crawl dữ liệu MXH tiếng Việt từ Youtube (200k bình luận)
  - Gán nhãn dữ liệu
- Xây dựng công cụ PTQĐ:
  - Tiền xử lý dữ liệu
  - Sinh vector biểu diễn từ, tính giá trị các đặc trưng
  - Chạy với các phương pháp phân loại
- Đánh giá trên dữ liệu tiếng Anh và so sánh với các hệ thống khác
- Lựa chọn mô hình tốt nhất áp dụng cho tiếng Việt

Kết quả cho tiếng Việt

	Accuracy (%)
DecisionTree	51.8
KNN	78.3
SVM	85.9
Voting	87.0

- Độ chính xác cao vì chỉ lấy dữ liệu từ 1 miền (âm nhạc)
- Tập dữ liệu test nhỏ so với tiếng Anh