**BÁO CÁO KẾT QUẢ THỰC NGHIỆM**

**1. Thực nghiệm và đánh giá**

**1.1. Dữ liệu thực nghiệm**

Tập dữ liệu D gồm 10661 câu bình luận (comments) Tiếng anh về lĩnh vực Phim., Từng dòng trong file rt-polarity.pos là 1 câu được gán nhãn tích cực, từng dòng trong file rt-polarity.neg là 1 câu được gán nhãn tiêu cực.

Số lượng đặc trưng dữ liệu câu bình luận (comments) Tiếng anh về lĩnh vực Phim.,. Biểu diễn thuộc tính của rt-polarity.pos và rt-polarity.neg theo bộ từ điển V thành các véc-tơ văn bản. File rt-polarity.pos có 5331 câu bình luận tích cực về phim. File rt-polarity.neg có 5330 câu bình luận tiêu cực về phim .Chúng tôi lựa chọn ngẫu nhiên 80% số lượng câu bình luận từ tập dữ liệu D sử dụng làm dữ liệu huấn luyện, còn lại 20% sử dụng làm dữ liệu kiểm thử.

**1.2. Môi trường thực nghiệm**

Công việc thực nghiệm được chúng tôi thực hiện trên máy tính cá nhân DELL với các thông số kĩ thuật như sau:

Processor Intel(R) CoreTM i5-7240M CPU @2.50 GHz, 12.00 GB memory.

Hệ điều hành Window 10 Pro

**1.3. Kết quả thực nghiệm**

Trong thực nghiệm, để khảo sát kết quả huấn luyện của mỗi loại hàm Ranking được lựa chọn, chúng tôi dụng phương pháp giảm chiều LDA để giảm chiều các véc tơ biểu diễn đặc trưng trong tập D và các độ đo Recall, Precision, và F1-score đánh giá hiệu quả của các mô hình học máy Phân lớp khác nhau: SVM, Randomforest, Logistic regression trên tập dữ liệu đã chia.

Dữ liệu giảm chiều phân tích ra kết quả kiểm thử dựa vào các độ đo và phương pháp giảm chiều LDA để đưa ra số chiều nhỏ nhất và các thuộc tính được phân loại đúng mà vẫn giữ nguyên được đúng tính chất không bị lẫn lộn. Dữ liệu sau khi phân tách thành 2 file các câu bình luận tích cực và tiêu cực về phim, không bị chồng lẫn lộn.

**1.4. Đánh giá kết quả**

Để so sánh kết quả huấn luyện và dự đoán hạng thuộc tính với các phương pháp khác chúng tôi cài đặt thêm thuật toán Randomforest, Logistic regression và sử dụng thuật toán hồi quy SVR (Support Vector) Regression) trong chương trình thực

Gọi Dtest là tập dữ liệu kiểm thử, theo D có 3 tiêu chí ảnh hưởng đến kết quả thực nghiệm được sử dụng đánh giá như sau:

**1.4.1**. Hàm mất mát:

Giả sử rằng chiều mà chúng ta muốn giảm về là D′<DD′<D và dữ liệu mới ứng với mỗi điểm dữ liệu xx là:

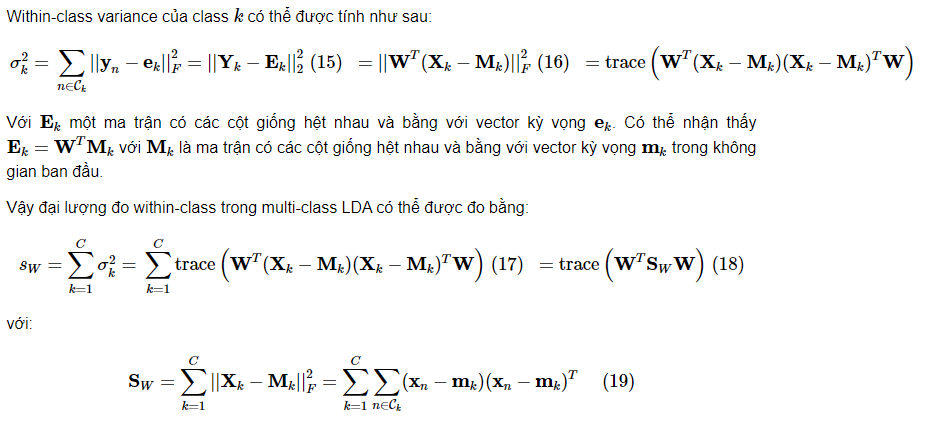


với *Chú ý rằng LDA ở đây không sử dụng bias.*

Một vài ký hiệu:

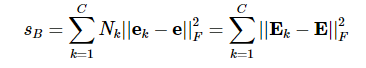
*  lần lượt là ma trận dữ liệu của class kk trong không gian ban đầu và không gian mới với số chiều nhỏ hơn.
*  là vector kỳ vọng của class kk trong không gian ban đầu.
* là vector kỳ vọng của class kk trong không gian mới.
* m là vector kỳ vọng của toàn bộ dữ liệu trong không gian ban đầu và e là vector kỳ vọng trong không gian mới.

**1.4.2.** Within-class nhỏ



**1.4.3**. Between-class lớn

Công thức tính:



Ta lấy Nk làm trọng số vì có thể có những class có nhiều phần tử so với các classes còn lại.Chú ý rằng ma trận E có thể có số cột *linh động*, phụ thuộc vào số cột của ma trận Ek mà nó đi cùng (và bằng Nk).



Với:

