BÀI TOÁN HỌC MÁY

***MSSV : 18120363 – Đặng Văn Hiển**

- **Mô tả bài toán :**
  - Task (T): Phân loại sắc thái bình luận của khách hàng vào 2 nhóm thái độ tích cực hay tiêu cực.
  - Performance measurement (P): Precision, Recall và F1 score.
  - Experience (E): tập hợp các bình luận đã được gán nhãn là tích cực hay tiêu cực.

- **Xây dựng mô hình học máy :**
  - Data: Dùng Tf-idf để thống kê tần xuất từ (n-gram) trong bình luận và mức độ quan trọng của nó. Rồi mô hình hoá không gian vector các đặc trưng là điểm Tf-idf của các từ trên (n-gram).

  - Hypothesis set: Tập các vector trọng số của đặc trưng tạo nên các siêu mặt phẳng chia không gian m chiều (m đặc trưng) với lời giải đích lý tưởng là một siêu mặt phẳng chia tách 2 tập điểm tích cực và tiêu cực trong không gian.

  - Algorithm: đề xuất thuật toán học sẽ sử dụng là gì? (chỉ nêu tên và mô tả chung)

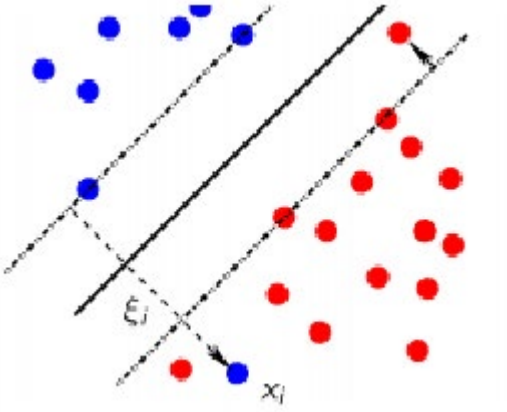
    *Thuật toán học được đề xuất là SVM (Support-vector machine)vì :

    - Vì các đặc trưng là từ được trích xuất bằng Tf-idf nên số lượng chiều của mỗi vector là rất lớn bao gồm mọi từ trong tập ngữ liệu. Trong những trường hợp này thì SVM phân loại hiệu quả hơn hẳn những thuật toán khác.

    - SVM giải quyết overfitting rất tốt (dữ liệu có nhiễu và tách dời nhóm hoặc dữ liệu huấn luyện quá ít).

- Phân lớp nhanh, có hiệu suất tổng hợp tốt và hiệu suất tính toán cao.

* Mô tả thuật toán SVM (Support-vector machine):



-Thuật toán SVM tìm siêu mặt phẳng chia đôi không gian (phân loại tích cực và tiêu cực) sao cho margin (khoảng cách cách đều đến 2 điểm gần nhất mỗi bên) là lớn nhất (và giảm số điểm bị phân loại sai – "soft" margin với tham số C hoặc dùng thủ thuật kernel) thì sai số tổng quát của SVM của thuật toán là nhỏ nhất.

$$\begin{aligned} \arg\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i \\ \text{subject to} \quad & y_i(\boldsymbol{w}\boldsymbol{x_i} + \boldsymbol{b}) \geq 1 - \xi_i \\ & \xi_i \geq 0, \quad i = 1, 2, \ldots, N \end{aligned}$$