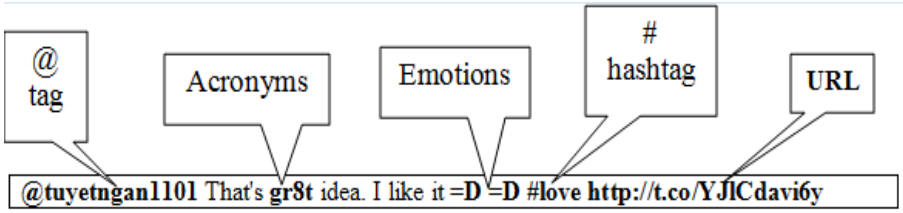

Xử lý tag, hashtag, URL:

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.