Facebook Posts Text Classification To improve Information Filtering

BENKHELIFA Randa, Laallam Fatima Zohra

randa.benkhelifa@univ_oargla.dz, Laallam.fatima-zohra@univ-ouargla.dz

Univ Ouargla, Faculté des Nouvelles Technologies de l'Information & de la Communication , Route de Ghardaïa, 30 000 Ouargla, Algérie.

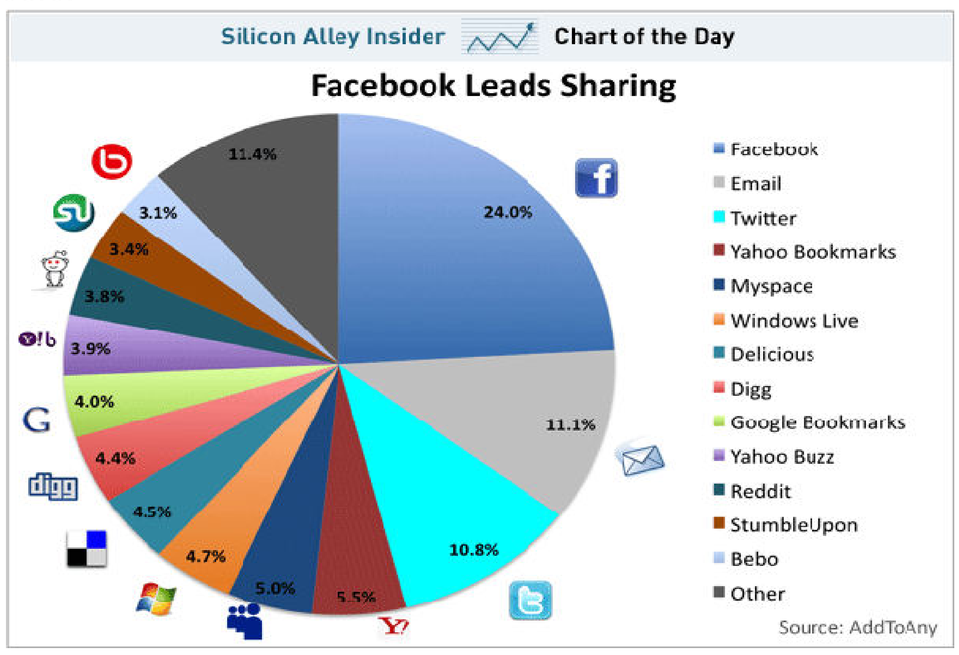
Abstract

Today, social networking websites are more than simple websites, they become a very popular tool of communication between users. They can easily access and share content, news, opinions and information in general. But until now Facebook rest by far the largest social networking website among Internet users in the world. Facebook statistics show that more than 1.2 billion users are active monthly. Average 21 minutes (time) spent per user per day. And Nearly 300,000 status updates are posted to Facebook every minute. In general, facebook user creates various text contents in the form of comments, wall posts, and blogs. This text has its own specific characteristics such as, massive data, noisy, dynamicity, contain unclear terms, and short. Usually Internet users do not care about the spellings and accurate grammatical construction of a sentence. They speak bravely using the slang terms (acronyms and abbreviations) in their posts. those posts disseminate information that is can be one of those six categories (news, opinion, deal, salutation, demand or quote). Almost previous works about social media text classification classify posts based on the topic (politic, sport, art, etc). In this work we have proposed to classify them in a swing level into the six pre- chosen classes using SVM, Naïve Bayes and K-NN machine learning methods We have also proposed a new approach to improving this classification showing the importance of StopWords and also the impact of internet slang on this classification.

Keywords. Facebook posts, Text Classification, Machine Learning, Internet Slang, Stopwords.

Introduction

Text pre-processing plays a major role in any categorization. Despite the impotence of this phase, its implementation is too difficult especially due to the nature (specific characteristics such as massive, dynamic, and noisy) of the text generated from social networks. Usually Internet users do not care about spelling and accurate grammatical construction of a sentence. They speak bravely using the slang terms (acronyms and abbreviations) in their posts. They employ several lexical units that appear syntactically different but in fact they describe the same meaning (E.g: $, dollars, dlrs,...). Facebook posts are perfect for these due to their abundance and short length. Moreover, Facebook is a popular social network with a great diversity of users.



The objectives of this work are:

- To classify posts in a lower level in order to filter some kind of posts.
- To study the problem of short text with special characteristics typically found on social media, and to show how much it is important to take these characteristics into consideration in the phase of pre-processing.
- To answer the questions: Are Stopwords always meaningless? Do they playing no role in improving classification regardless the categories?

Methodology

Dataset

20000 recent Facebook' text posts in six predefined categories: 4200 News, 6600 Opinions, 3000 Quotes, 3000 Deals, 1500 Salutations and 1700 Demands

Data Preprocessing

Algorithms

- **Algo1** — Detect and correct Internet Slang terms

L8R → Algo1 → Later

2day → Algo1 → Today

- **Algo2** — Detect and replace Internet Slang terms by the same lexical unite

L8R, 2day, ... → Algo2 → SI

- **Algo3** — Detect and replace accuracy and percentage terms by the same lexical unite

dollars, 50$, 25£ → Algo3 → $

Preprocessing

- ✓A term that appears less than 3 times is removed;
- ✓Removing punctuation (.,!?) and symbols ([<>));
- ✓Using loven_Stemmer;
- ✓Use TF-IDF as features' selection.

Proposed Approach

The alternative features token in each pre-processing method (P1-P6).

	P1	P2	P3	P4	P5	P6
Removing Stopwords	✓	✗	✗	✗	✗	✗
Algo1	✗	✗	✓	✗	✗	✗
Algo2	✗	✗	✗	✓	✗	✓
Algo3	✗	✗	✗	✗	✓	✓

Tab 1: Proposed Approaches

Classifiers

- K-NN
- Naïve Bayes
- SVM

Results

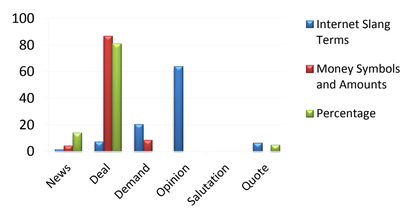


Figure 1: Distribution of terms in each class

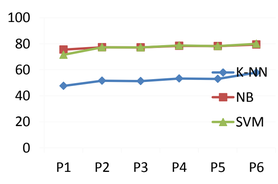


Figure 2: F-Measure results

	P1	P2	P3	P4	P5	P6
K-NN	52.2	55.57	56.47	57.3	56.72	61.3
NB	75.5	76.86	76.11	78.4	77.26	79.4
SVM	71.7	77.31	77.26	78.9	78.21	80.3

The best result is gotten by SVM

Tab 5: Accuracy Results.

Conclusion

This work is aimed at classifying Facebook text posts according to a new set of selected categories We conclude that Stopwords are not always meaningless, as they play a major role in improving the performance of some classification, depending on the category at hand. Also Internet slang terms, which contribute to improve the classification performance.