Social Post Analysis Using Naïve Bayes and Core NLP

Abhijit Bhoske, Mukul Anaspure, Ashis Bhoi, Pankaj Karkhele, A.S.Deokar

Abstract: In today’s world, daily there is enormous information published on the web (social media like facebook, twitter, etc). This information contains movie reviews, product reviews, blogs, news articles, etc. It is not easy to predict the kind of information to which it belongs. So in order to solve the above-mentioned issue, we have proposed the system in which when any post that contains textual information given as an input, our system tries to provide a solution from the web. To make a useful post for business, the system extracts useful information from the text. The use of the system is to take a post directly to its potential audience (online users on social media). Here, the proposed system analyses the social media posts and understands what kind of decisions the user may take in the future. So here the proposed system can recommend a solution to the user directly with a certain post. There are certain domains which we will identify from the post. Content will be suggested from the post to the potential audience (business users) and potential audience (business users) will recommend the solution or suggestion to the user.

Keywords: Core NLP, Keyword Extract, Regular Expression, Entity Extraction

Introduction:

Each and every day, there are lots and lots of contents being published on the web. After some days the post becomes useless. So we are developing a system. Our proposed system is going to be useful for social media where textual information is posted. If a blog post on a site isn’t viewed by the appropriate audience, then the number of online business users on the site are useless and the sales may also be low. On the other hand, if an educational article, which is rich in content, doesn’t reach many people, then a student seeking knowledge under that particular topic will lose a good source of knowledge. This is going to be a losing situation for both content providers and content seekers.

Related work:

There are some researches that predict the popularity of the web content. Volume of online post, news stories would receive an Predicting textual, semantic, real-world, surface and cumulative features. Textual features denote certain discriminative terms like ‘India’, ‘Mobile’, ‘Friend’, etc., from each different news sources. Semantic feature denotes named entities such as locations, people, organizations, etc. Real-world features are the correlations between environmental conditions like weather conditions and commenting behaviors. Meta features like quality of the post sources and news agents are represented as surface features and the number of times a particular post is published by various news agents is denoted as the cumulative features. All these studies are about predicting the popularity of acontent, but ours is mainly to derive rules to propose changes to be done to a post in order to provide a solution on the web. Also most of the features used in these studies are mainly numerical measures. But we have exploited in our study some subjective elements like emotions and sentiments. We have also incorporated intention mining in our study. Intention Mining is a novice subject area which is at its early stages of development. In our study, intention mining is used to predict whether a person is likely to watch a movie or not.

Motivation:

There are no any existing systems which work on any kind of operation on the post.

After posting content on social media, most of the posts after a certain period become useless.
So the prime focus is to make those posts useful.

Our system can provide the solution to related post.

The system can generate advertising for a business user.

According to the need of social users, the business users can suggest the solution.

**System Architecture:**

In this user can post text as an input. Using core NLP techniques, given text file or code file will be processed. Proposed systems going to perform operations like stemming, stop words removal and parsing techniques.

**Core NLP Technique:**

Tokenization – the process of converting a text into tokens.

Stemming: Stemming is a rudimentary rule-based process of stripping the suffixes (“ing”, “ly”, “es”, “s”, etc) from a word.

Stop word removal: Language stop words (commonly used words of a language – is, am, the, of, in etc.), URLs or links, social media entities (mentions, hash tags), punctuations and industry-specific words. This step deals with removal of all types of noisy entities present in the text.

Entity Extraction: Entities are defined as the most important chunks of a sentence – noun phrases, verb phrases or both. Entity Detection algorithms are generally ensemble models of rule-based parsing, dictionary lookups, post tagging, and dependency parsing. The applicability of entity detection can be seen in the automated chat bots, content analyzers, and consumer insight.

**Conclusion:**

Proposed system analyzes the post of a social user. It understands his/her problem as well as a requirement like watching a movie in the future, shopping, education, etc. Notification to the related business user is generated if the notification is related to that business user. The business user makes advertise for the social user and notifies them.

**Advantages:**

- Access to authorized personnel only.
- User-friendly.
- Memory space utilized efficiently.
- Multiple algorithms working together to produce best results.

**Disadvantages:**

- May give variable accuracy.
- Our system will work only on textual content not on images.
- Our system will work on the basis of probability.
References:


[12] A retraction: papers that plagiarize only text can still contribute to the literature, but any errors or omissions should be prominently corrected, says Praveen Chaddah.” Nature 511.7508 (2014): 127-128.

