Viral Post Identification using Core NLP and Naïve Bayes

Kalyani Sonawane, Sudha Kulkarni, Gauri Dhamale, Anuja Yedle, Prof. Nilam I. Dalavi

Abstract: In the today’s world every day there is enormous information is published on the web (social media, science and more). This information contains movie reviews, product reviews, blogs, news articles, etc. It is not easy to predict this kind information to which it belongs. So proposed system need to solve the above-mentioned issue for that we proposed the system in which when any post that contains textual information given as an input, makes it to provide solution from the web. To make a post for business the system extract useful information from the text. The use of the system is to take a post directly to its potential audience (online users like social media). Here, proposed system analyze the social media posts and understand what kind of decisions they may take in the future so that proposed system can recommend to the user directly with a certain post. There are certain domains which we will identify from the post. Content will suggest from the post to the potential audience and potential audience 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 of the days the post is useless so that we are developing a system. Our proposed system is going to be useful for social media where textual information is post. If a blog post on a site doesn’t get viewed by the appropriate audience, then the number of online business users on the site are useless and the sales may be low On the other hand, if an educational article, which is rich in content, doesn’t reach many, then a student seeking knowledge under that particular topic will lose a good source of knowledge. This is going to be aloe situation for both content providers and content seekers.

Fig. basic Structure

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 a content, but ours is mainly to derive rules to propose changes to be done to a post in order to make it provide 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 not any existing system who work any kind of operation on the post

After posting content on social media some of the time posts is useless

So the prime focus is to make those post as useful
Our system can provide the solution to related post

The system can generate advertising for a business user

Accordingly, the need of social users business user can suggest the solution

**System Architecture:**

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

![Fig 1 system architecture](image)

**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 analyze the posts of a social user. Understand his problem as well as a requirement like watching a movie in the future, shopping, education, etc. Notify to the related business user if the generated notification is related to the business user the business user make advertise for the social user and notify them

**Advantage:**

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

**Disadvantage:**

- May give variable accuracy.
- Our system will work only on textual content not on image.
- Our system will work on 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.

