Viral Post Identification using Core NLP Technique

Kalyani Sonawane
Gauri Dhamale
Sudha Kulkarni
Anuja Yedle

Prof. Nilam I. Dalvi

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 fromthe 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 [9]. Content will suggest from the post to the potential audience and potential audience will recommend the solution or suggestion to the user.

IndexTerms **- Component,formatting,style,styling,insert.**

I. 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 [5]. 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 aloes situation for both content providers and content seekers [11].

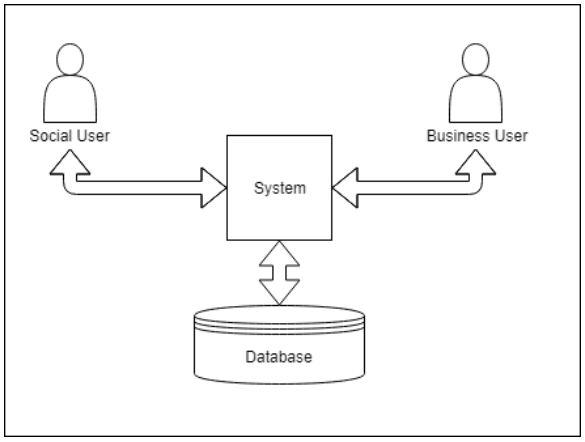

Fig. basic Structure

Keywords

CoreNLP, Keyword Extract, Regular Expression,Entity Extraction

Related Work:

There are some researches that predict the popularity of theweb content. Volume of online post, news stories would receive anpredicting 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 aparticular post is published by various news agents is denoted
as the cumulative features [4]. All these studies are about predicting the popularity of content, but ours is mainly to derive rules to propose changesto 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 mainlynumerical measures. But we have exploited in our study somesubjective elements like emotions and sentiments. We havealso incorporated intention mining in our study. IntentionMining 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 postafter posting content on social media some of the time posts is uselessso the prime focus is to make those post as usefulOur system can provide the solution to related postThe system can generate advertising for a business user
Accordingly, the need of social users,business user can suggest the solution.

Mathematical Model

Let the proposed system be defined by set theory as:

Input: Posted Text

Output: Solution related to post

S = {s, e, X, Y}

s = Start of the program

1. Register/Login into the system

2. Text posted by user

e = End of the program

X = input of the program

= {I}

I = Text

Y = Output of program = Solution by business user

First, user will post text on the system that will contain some information.

System extracts features with the help of Naïve Bayes and core NLP.

Let F be the set of features

F= {F1, F2 ..., Fn}

These features are compared with extracted features of training dataset. The classifier classifies these features and gives solution to the user

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 systemare going to perform operation like stemming, stop words removal and parsing technique.

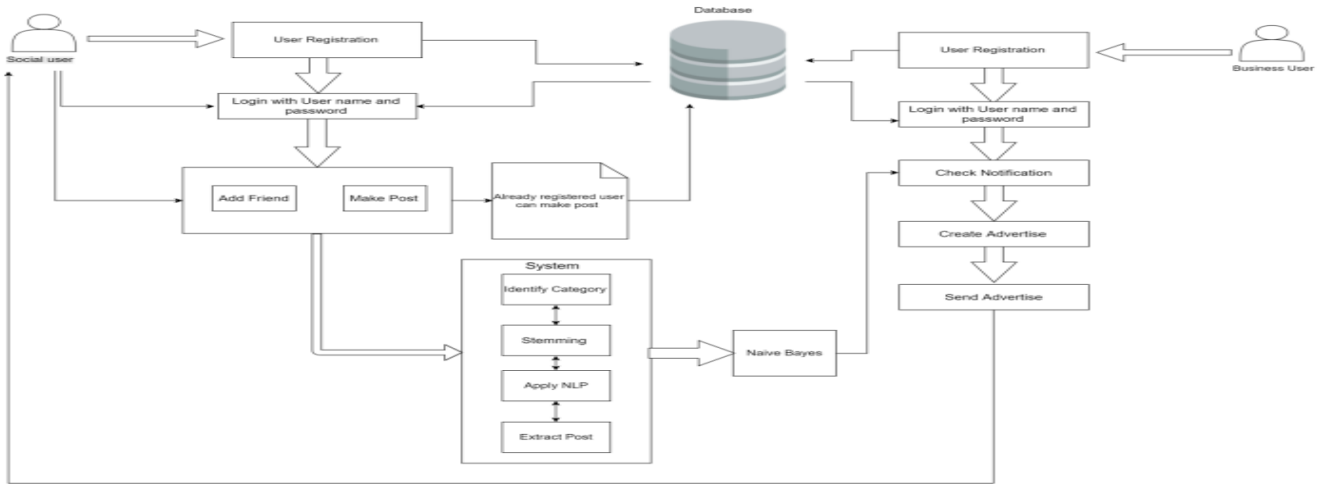

Fig.1 System Architecture

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 [14]. 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 analysers, and consumer insight.

Implementation steps:

- **Social User**

  Step 1: RegisterUser

Step 2: User Login

  Step 3: Create Post

  Step 4: Upload Post

- **Business User**

  Step 1: RegisterBusiness User

  Step 2: Business User Login

  Step 3: See Social user's post

  Step 4: Create Advertise Related to Post

  Step 5: Send Notification to Social User

Result:

System provide a solution on the real-time post. A business user can make a business from the post. The system will generate a summary of the enormous post and sends a summary to the related business user. Find the potential business user of the post and suggest them the post directly. The first part is the generations of rules that make a post go on social media, business user analyzes an actual post and give suggestions on making the post go viral.

Conclusion:

Proposed system analyzes 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.

ACKNOWLEDGMENT

It gives us great pleasure in presenting the preliminary project report on '**Viral Post Identification using Core NLP and Naïve Bayes'**.

I would like to take this opportunity to thank my internal guide for giving me all the help and guidance I needed I am really grateful to them for their kind support. Their valuable suggestions were very helpful. I am also grateful to Head of Computer Engineering Department, for his indispensable support and suggestions.

Name of Students
Kalyani Sonawane, Gauri Dhamale, Sudha kulkarni, Anuja yedle

REFERENCES

[I] L. Kong, Z. Lu, H. Qi, and Z. Han, "Detecting High Obfuscation Plagiarism: Exploring Multi-Features Fusion via Machine Learning," 1nl. 1. u-and e-Service. Sci. Technol., vo!. 7, no. 4, pp. 385-396, 2014.

[2] Izzat Alsmadi1, Ikdam AlHami2 and Saif Kazakzeh3 " **Issues Related to the Detection of Source Code Plagiarism in Students Assignments**" International Journal of Software Engineering and Its Applications Vol.8, No.4 (2014), pp.23-34 http://dx.doi.org/10.14257/ijseia.2014.8.4.03

[3]M. K. Shenoy, K. C. Shet, and U. D. Acharya, "Semantic Plagiarism Detection System Using Ontology Mapping," Adv. Compul. An Inl. 1., vol. 3, no. 3, pp. 59-62, May 2012.

[4] S. Alzahrani and N. Salim, "Fuzzy semantic-based string similarity for extrinsic plagiarism detection," Braschler and Harman, 2010.

[5] S. Harispe, D. Simchez, S. Ranwez, S. Janaqi, and J. Montmain, "A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain," 1. Biomed. In{iirm., vo!. 48, pp. 38-53, Apr 2014.

[6] J. O. Shea, Z. Bandar, K. Crockett, and D. Mclean, "A Comparative Study of Two Short Text Semantic Similarity Measures," Arlij: Inlell., vo!. 4953, pp. 172-181,2008.
[7] K. Bazdaric, V. Pupovac, L. Bili6-Zulle, and M. Petrovecki, "Plagiarism as a violation of scientific and academic integrity," 2009.

[8] E. S. AI-Shamery and H. Q. Gheni, "Plagiarism Detection using Semantic Analysis," Indian 1. Sci. Technol., vol. 9, no. I, Feb. 2016.

[9] R. Mihalcea, C. Corley, and C. Strapparava, "Corpus-based and knowledge-based measures of text semantic similarity," 2006, vo!. 6, pp. 775-780.

[10]Agrawal, Mayank, and Dilip Kumar Sharma. "A state of art on source code plagiarism detection." *Next Generation Computing Technologies (NGCT), 2016 2nd International Conference on*. IEEE, 2016

[11]Ragkhitwetsagul, Chaiyong, and Jens Krinke. "Using compilation/decompilation to enhance clone detection." *Software Clones (IWSC), 2017 IEEE 11th International Workshop on*. IEEE, 2017

[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.

[13]Oberreuter, Gabriel, and Juan D. VeláSquez. "Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style." *Expert Systems with Applications* 40.9 (2013): 3756-3763.

[14] Nilsson, Lars-Erik, Anders Eklöf,and Tina Kullenberg. "Categorizing students, categorizing texts: will plagiarism detection leave blood on the tracks?." *Earli 2017 Biennal conference*. 2017.

[15]https://en.wikipedia.org/wiki/Plagiarism