International Journal of Future Generation Communication and Networking

Vol. 13, No. 3s, (2020), pp. 805–810

Social Post Analysis Using Nlp Techniques

Omkar Joshi[#1], Abhijeet Pawar[*2], Anmol Goja[#3], Hrushikesh Navgire[*4], Balram Somani[#5]

#Dept of Information Technology, Sinhgad Institute of Technology and Science

[1]omkar_21@yahoo.com

[2]Abhijeetpawar7353@gmail.com

[3]Agoja3@gmail.com

[4]hrushikeshnavgire@gmail.com

[5]balramsomani35@gmail.com

Abstract

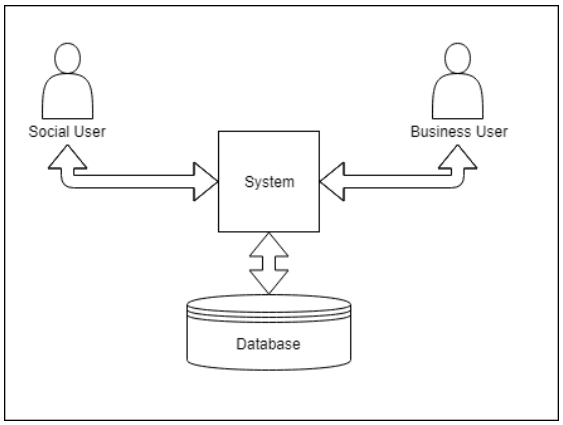
In the today's world every day there is enormous information is published on the web (social media, science and more). This information contains movie reviews, product reviews, blogs, news articles, etc. It is not easy to predict this kind information to which it belongs. So proposed system needs to 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 extracts 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 analyses 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.

Keyword: CoreNLP, Keyword Extract, Regular Expression, Entity Extraction

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. 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.

ISSN: 2233-7857 IJFGCN



There are some researches that predict the popularity of the web content. volume of online post, news stories would receive a predicting textual, semantic, real-world, surface and cumulative feature. 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 behaviours. 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.

II. LITERATURE SURVEY

In recent years, there have been some researchers focused on application of text mining on their research. In 2015, Gulo and Rubi, in their study on developing text mining solution for scientific articles using R language found out that simple depiction of documents was as effective as depictions connecting more complicated analysis.

III. METHODOLOGY

Our System mainly contains two modules one is Social media and another is Business User.

Social Media:

In this Module we are creating a social media like Facebook or twitter. we are added a functionality like Adding profile picture, posting tweets, adding friends and removing friend from there profile. Social Media user can view the posts of other users.

ISSN: 2233-7857 IJFGCN

Business User:

In this module Business user can suggest or make Advertisement of their project directly to the Social user for their Business by Notification.

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

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 analyser's, and consumer insight.

IV. 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}

ISSN: 2233-7857 IJFGCN

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

ALGORITHM OF EXTRACTION:

1. TF-IDF

2. Latent Dirichlet Allocation (LDA)

3. Naive Bayesian Model

TF-IDF (Term Frequency-Inverse Document Frequency):

IDF algorithm is used for analysing the data and calculating the text frequency. Text frequency, it is no. of times a text as appeared in an article. For example, in a newspaper sports article how many times the word cricket has appeared. TF-IDF algorithm will be used for analysing user's post.TF-IDF algorithm will indicate how many times same word is occurred in all the post which are uploaded by users. Calculating text frequency will help business user to analyse and improve their business from the current trend.

LATENT DIRICHLET ALLOCATION (LDA):

Latent Dirichlet Allocation (LDA) model for efficient discovery of suspicious social assorting groups in tourism domain. Latent Dirichlet Allocation (LDA) is a generative statistical model that allows sets of observations to be explained by observing groups that explain why some parts of the data are similar.

NAÏVE BAYESIAN MODEL:

Naive Bayesian Model is an algorithm which uses data and gives output as classified text. It analyses the text in the social media related to the customer health terminology. Naive Bayesian algorithm comes under supervised learning in machine learning. supervised learning works with the help of labelled output. Naive Bayesian classifies the text and find the probability of the outcome. For example, based on previous data of weather, it will predict form the probability of a person can play outside in this weather or not.

ISSN: 2233-7857 IJFGCN

V. PROPOSED SYSTEM

The proposed system implements social analyser. It shows relationship between different components of our system, it shows how business user and social user interact in our system using machine learning and NLP techniques. It shows how data is stored in our system. How post is classified using naive bias for the further process of social analyser.

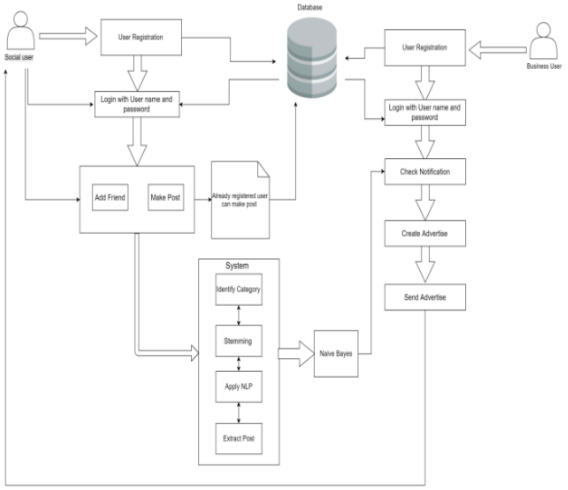


Fig block diagram of proposed system

VI. IMPLEMENTATION AND RESULT

SOFTWARE VERIFICATION:
Evaluating the component to determine whether the product of given development phase satisfy the condition imposed at the start of the development.
•The buttons are in working conditions is checked (e.g. Home page).
•Walk through is performed with current modules.
•The errors are noted and corrected in verification.

SOFTWARE VALIDATION:
In this testing process, the evaluation of components is done by actually executing the code and checked whether the requirements are fulfilled or not.
•The validation is done by executing the code of developed modules of user registration and login page, home screen, text plagiarism, code plagiarism, text and code plagiarism page.
•Black box and White box testing of modules are done.
•The technical errors are noted and corrected.

VII. CONCLUSION

We 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. Our system sends a summary to the related business user. Find the potential business user of the post and suggest them the post directly. The first

ISSN: 2233-7857 IJFGCN

part is the generation of rules that make a post go on social media and we analyse an actual post and give suggestions on making the post go viral.

REFERENCES

[1] Liu.B (2012), Sentiment Analysis and Opinion Mining, Morgan amp; Claypool Publishers.

[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).

[3] TsagkiasWeerkamp/5c6de157e19b49ff007f24c04c1f24d91addb6ba M. Taboada et al (2010), Lexicon-Based Methods for Sentiment Analysis, In proceedings of the Annual Meeting of the Association for Computational Linguistics, Singapore 3.association-for-computational-linguistics-volume-1-long-papers Miller and Charles (2016), A psychological based analysis of Marketing Email Subject Lines, In proceedings of the International Conference on Advances in ICT for Emerging Regions ICTer2016.

[4] Omkar Sunil Joshi and Garry Simon, "Sentiment Analysis Tool on Cloud: Software as a Service Model,"2018 International Conference on Advances in Communication and Computing Technology (ICACCT), *Sangamner, 2018, pp. 459-462*.

[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., vol 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., vol. 4953, pp. 172-181,2008.

[7] Omkar Sunil Joshi, Bhargavi R. Upadhvay and M. Supriya, "Parallelized Advanced Rabin-Karp Algorithm for String Matching," *2017* International Conference on Computing, Communication, Control and Automation (ICCUBEA), Pune, 2017, pp. 1-5.

[8] Vinodhini G. and Chandrasekaran.R (2012), Sentiment Analysis and Opinion Mining: A Survey, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 6.

ISSN: 2233-7857 IJFGCN