iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** IV **Month of publication:** April 2018

DOI: http://doi.org/10.22214/ijraset.2018.4482

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

Topic Detection IN Twitter with Data Mining

Dhruvika A. Mistry[1], Aakash Shah[2]
[1, 2]*Computer engineering, Silver oak college of engineering & technology, Gujarat Technological University*

Abstract: Twitter is a user-generated content system that allows its users to share short text messages with 140 characters which called Tweets. Detecting topics from the large amount of tweets can help to facilitate many downstream application of intelligent computing. Detecting topic from large twitter data can be used for natural disaster warning, user opinion assessment and traffic prediction etc. For detecting coherent topics from tweeter dataset, many techniques are proposed most of which are using LDA with some modifications. LDA is more ML (machine learning), tweeter dataset is more language oriented. So we have used NLP for pre-processing, and sentiment analysis. Performing LDA will be more time consuming then sentiment analysis.
Keywords: Twitter, Data mining, Topic Detection, text mining, Sentiment analysis

I. INTRODUCTION

Social media generates a large amount of data and information. In this information may contain many topics that can be useful for many downstream applications. There is large number of twitter users around 695 million in 2016 which generates 58 million tweets daily. With this large amount of data user can miss important topic. It is big task to mine coherent topic from user generated, unstructured data.

Detecting topic can be useful in many things like discovering natural disaster as early as possible, helping political parties, box-office prediction for movies, companies to understand user's opinions, improving content marketing by better understanding customer needs, understanding students problems related to current learning method etc. Mining topics from fast evolving data in social media would improve the performance of many downstream applications [1]. Traditional topic models work on large size of documents. Traditional topic models need some specific length of documents to perform topic detection. It is difficult to perform this traditional models on social media text, specifically on tweets as tweets are short text message of 140(now 280) characters. One key weakness of topic models is that they need documents with certain length to provide reliable statistics for generating coherent topics. In Twitter, the users' tweets are mostly short and noisy. Observations of word co-occurrences are incomprehensible for topic models. To deal with this problem, previous work tried to incorporate prior knowledge to obtain better results. However, this strategy is not practical for the fast evolving user generated content in Twitter.

From studying the research papers for literature review found that most of technique used machine learning as they have used LDA. But with machine learning the process becomes lengthy, less efficient, and more complex as these are performed on tweets (text), it is more part of text mining to detect topic from tweets. So, we have used sentiment analysis for detecting words having some meaning for further process. Previously to find coherent topics from twitter a SILDA model [1] is introduced which uses pre-learnt interest knowledge to find coherent topics from twitter. For that they have first do community detection and then perform LDA on sub data set to generate interest word set which will later applied to SILDA process. In this paper, we have used sentiment analysis for finding interest word set. To achieve better performance our result parameters will be accuracy precision and recall.

II. BACKGROUND

A. Text Mining

Text mining, also referred to as text data mining, roughly equivalent to text analytics, is the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities). Text analysis involves information retrieval, lexical analysis to study word frequency distributions, pattern recognition, tagging/annotation, information extraction, data mining techniques including link and association analysis, visualization, and predictive analytics. The overarching goal is, essentially, to turn text into data for analysis, via application of natural language processing (NLP) and analytical methods. A typical application is to scan a set of

documents written in a natural language and either model the document set for predictive classification purposes or populate a database or search index with the information extracted.

B. Natural Language Processing(NLP)

Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken. NLP is a component of artificial intelligence (AI).NLP can be used to interpret free text and make it analysable. Google and other search engines base their machine translation technology on NLP deep learning models. This allows algorithms to read text on a webpage, interpret its meaning and translate it to another language. Prior to deep learning-based NLP models, this information stored in free text files was inaccessible to computer-assisted analysis and could not be analyzed in any kind of systematic way. But NLP allows analysts to sift through massive troves of free text to find relevant information in the files. Current approaches to NLP are based on deep learning, a type of AI that examines and uses patterns in data to improve a program's understanding. Deep learning models require massive amounts of labelled data to train on and identify relevant correlations, and assembling this kind of big data set is one of the main hurdles to NLP currently. Earlier approaches to NLP involved a more rules-based approach, where simpler machine learning algorithms were told what words and phrases to look for in text and given specific responses when those phrases appeared. But deep learning is a more flexible, intuitive approach in which algorithms learn to identify speakers' intent from many examples, almost like how a child would learn human language.

C. Algorithm for NLP

Here are some algorithms that used for the NLP.

1) For Chunking, Named Entity Extraction, POS Tagging: - CRF++, HMM
2) Word Alignment in Machine translation: - Maxent
3) Spell Checker: - Edit Distance, Soundex
4) Parsing: - CKY algorithm and other chart parsing algorithms
5) Document Classification: - SVM, Navie bayes
6) Anaphora Resolution: - Hobbs Algo, Lippin and Leass algo, Centering Theory
7) Topic Modeling and keyword extraction: - LDA, LSI

D. Sentiment Analysis

The process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writers' attitude towards the particular topic product etc. is positive, negative or neutral. Sentiment analysis is the measurement of positive and negative language. Sentiment analysis is extremely useful in social media monitoring as it allows us to gain an overview of the wider public behind certain topics.

E. Community Detection

In the study of complex networks, a network is said to have community structure if the nodes of the network can be easily grouped into (potentially overlapping) sets of nodes such that each set of nodes is densely connected internally. In the particular case of non-overlapping community finding, this implies that the network divides naturally into groups of nodes with dense connections internally and sparser connections between groups. But overlapping communities are also allowed. The more general definition is based on the principle that pairs of nodes are more likely to be connected if they are both members of the same community(ies), and less likely to be connected if they do not share communities. A related but different problem is community search here the goal is to find a community that a certain vertex belongs to.

III.RELATED WORK

We have referred [1] for topic detection method. They have introduced topic model called SILDA. This model uses the pre-learnt interest knowledge for topic detection. For that first they have clustered the users according to the retweeting network using community detection and generate data subsets. The users' interests are mined as the prior knowledge. Interest sets are generated by performing LDA on the subsets. That data is then applied to improve the performance of topic learning. They have proposed to learn prior knowledge from data itself. The traditional topic model performs poorly because of the noisy and short tweets. In SILDA the data set is divided into small less noisy sub datasets. Text in same sub dataset shares similar topics. The partition of dataset is done on the base of the retweeting behaviour of users. For partition here they have used community detection algorithm SLM. And then LDA is performed on the tweets in ach community. Distinctive topics are minded as the common interest of the corresponding users [1]. This sub dataset will then generated interest word set by performing LDA on the sub dataset. This interest word set will be used

further for topic detection as prior knowledge. They have applied two metrics for the performance of the model. Topic coherence is proposed to evaluate the quality of topics [1]. And Jaccard coefficient is used to measuring the similarity between to finite dataset [1].

IV. SILDA WITH SENTIMENT ANALYSIS

Most of the previous topic detection algorithms have used the LDA algorithm, and with some modification topic detection is done. But LDA uses a generative approach for topic detection so this process becomes more time consuming. NLP (Natural Processing Language) is used for pre-processing of data sets and then community detection is done. After that sentiment analysis is done on the community sets, after that the step 3 of SILDA [1] process are performed. This will get a list of coherent topics.

This process will get the coherent topic from the Target topic or the given dataset. For the detection of coherent topic the parameters will be accuracy, precision and recall.

For pre-processing NLP tasks (tokenization, POS tagging etc.) are performed using NLTK package, Gensim package is used for topic modelling and document similarity.

V. PROCESS FOR COHERENT TOPIC DETECTION

A. *Dataset*

We can get tweeter dataset from (https://www.kaggle.com/noahgift/social-power-nba) or we can also get tweet data set from twitter API, for that we have to install tweepy. Tweepy is open sourced library hosted on GitHub and enables Python to communicate with Twitter platform and use its API to search tweet dataset.

To get access on tweepy library we need access_token, access_token_secret, consumer_key, and consumer_secret. This program will return unstructured data for specific time period which we entered in the program. And then to get the data in human understandable format we have to save that unstructured data in JSON (JavaScript Object Notation) format.

B. *Pre-processing*

Next step will be pre-processing, where tokenization, stopping and stemming will performed. In the tokenization step the data will divide into tokens, then stopping will remove stop words and stemming will converts words into verb form.

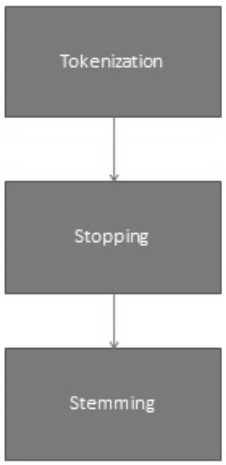


Fig. 1 Pre-processing steps

C. *Proposed flow(algorithm)*

Step 1: gather twitter dataset (any coherent topic selected- for exa : donald trump)
Step 2: read csv using textBlob (python) or download tweets (using twitter api)
Step 3: pre-processing of words / sentences

1) Tokenization (nltk)
2) Stopping
3) Stemming

Step 4: SLM for community detection

Step 5: perform sentiment analysis and store in a vector.

(In addition, this algorithm provides a Sentiment By Term algorithm, which analyzes a document, and tries to find the sentiment for the given set of terms. The algorithm works by taking in a string, a list of terms, and then splits the document into sentences, and computes the average sentiment of each term. This algorithm becomes powerful when combined with an auto-tagging algorithm, such as LDA, Auto-Tag URL, or Named Entity Recognition algorithms.)

Step 5 (step 3 of SILDA)

a) Topic T ← Topic(1,n)

b) topic t ← T(0,n)

c) interest I ← S(1, i)

d) Emit word wd,t ← T(t(0,n), S(1,i))

Where T is dataset topic(target topic),

t is topic from pre-processed dataset topic list,

I is list of words after sentiment analysis,

wd is coherent topic(word)

Step 7: results will be displayed

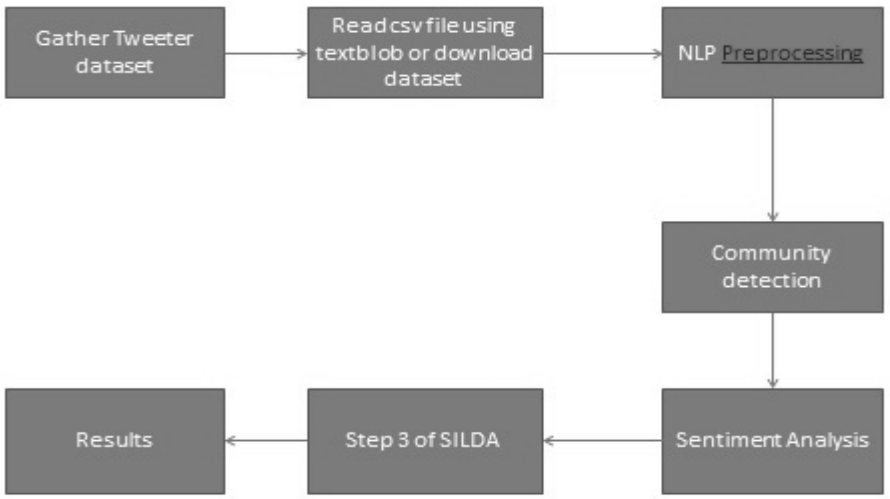


Fig. 2 proposed flow.

VI. CONCLUSION

We have used the generalized tweeter dataset for the coherent topic detection in tweeter. To get the topics first we have gathered the dataset and did pre-processing on it, then we have done the sentiment analysis on the dataset which we created by performing community detection. After performing sentiment analysis we get the word set in a file. And then we perform the step 3 of SILDA. This will get result list of coherent topics. This topic list we get on parameter of accuracy, precision and recall. In this method we use NLP processing, LDA will consume more time on finding topics. So we have used sentiment analysis from which we can get the writers attitude towards the target topic is positive or negative, and the word list which is based on the sentiments of words. Till now we get the accuracy. Now the remaining work is to get output in format of coherent topic list and find out the precision and recall parameter results.

REFERENCES

[1] Yuan He, Cheng Wang, Senior Member, IEEE, Changjun Jiangz "Mining Coherent Topics with Pre-learned Interest Knowledge" in Twitter IEEE Access,2017, volume 5, IEEE journoals & magazines, page: 10515-10525

[2] Rania Ibrahim, Ahmed Elbagoury, Mohamed S. Kamel, Fakhri Karray "Tools and approaches for topic detection from Twitter Streams: survey" Springer july2017 Knowledge and information system march 2018, volume 54, issue 3, pp 511-539

[3] Luca Maria Aiello, Georgios Petkos, Carlos Martin, David Corney, Symeon apadopoulos, Ryan Skraba, Ayse , Yiannis Kompatsiaris, Alejandro Jaimes "Sensing trending topics in Twitter" IEEE transaction on Multimedia, 2013, volume 15, issue 6, page 1268-1282

[4] Ahmed Elbagoury,_ Rania Ibrahim,_Ahmed K. Farahat, Mohamed S. Kamel, and Fakhri Karray "Exemplar-Based Topic Detection in Twitter Streams", Conference: International AAAI Conference on Web and Social Media – 2015

[5] David M. Blei ,Andrew Y. Ng ANG, Michael I. Jordan JORDAN "Latent Dirichlet Allocation" Editor: John Lafferty ,Part of: Advances in Neural Information Processing Systems 23 (NIPS 2010)

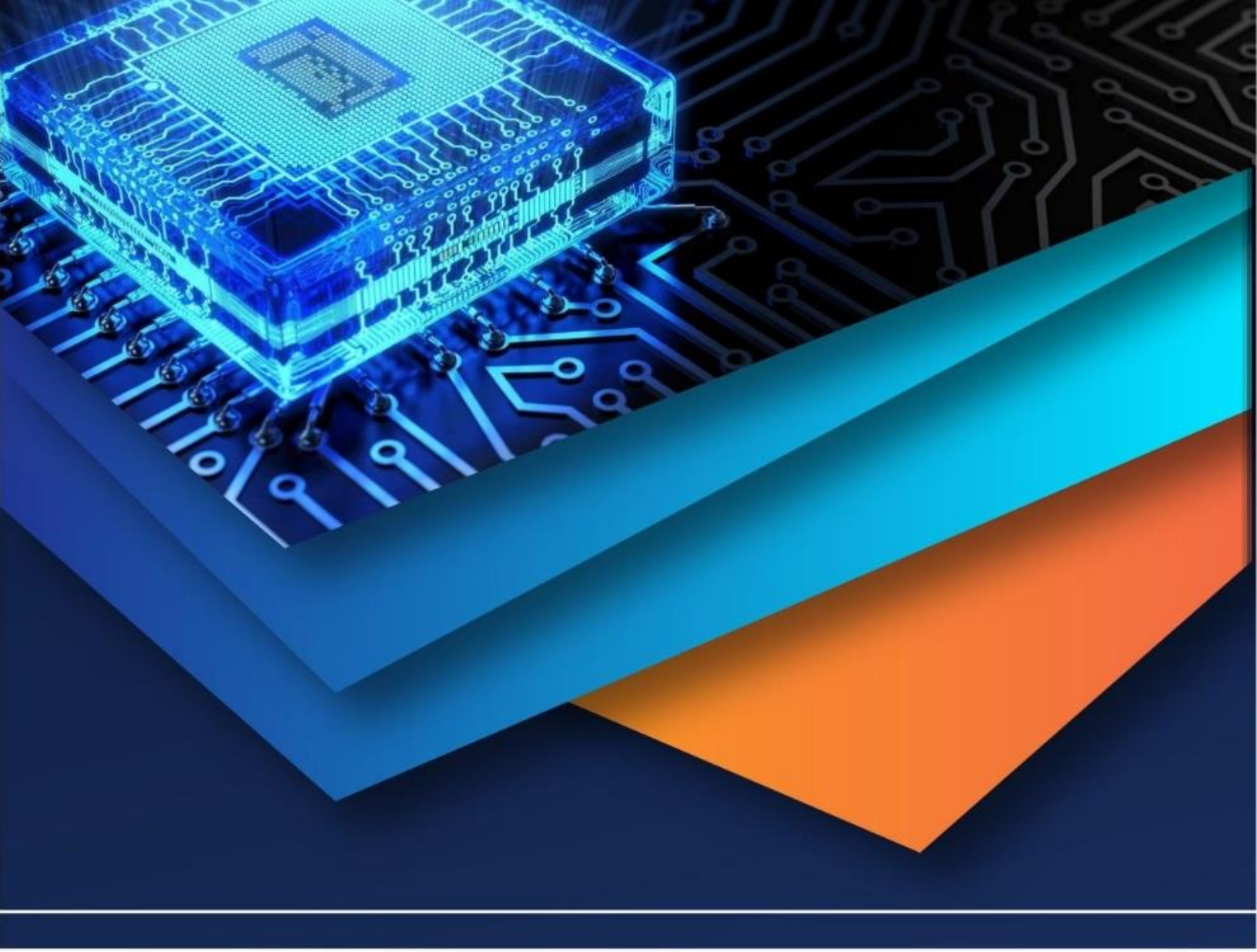
[6] Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, Riddhiman Ghosh "Leveraging Multi-Domain Prior Knowledge in Topic Models", IJCAI '13 Proceedings of the Twenty-Third international joint conference on Artificial Intelligence

[7] Wayne Xin ZHAO, Ee Peng, LIM,Jing JIANG "Comparing Twitter and Traditional Media using Topic Models", Part of the Lecture Notes in Computer Science book series (LNCS, volume 6611)

[8] Ludo Waltman and Nees Jan van Eck "A smart local moving algorithm for large-scale modularity-based community" detection The European Physical Journal B November 2013, 86:471

[9] Rajani D.Gavali, Prof. A.R.Kulkarni "Summarization of Social Media Data Using Topic Detection", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 2, February 2017

[10] Saif M. Mohammad1, Svetlana Kiritchenko1, Parinaz Sobhani2, Xiaodan Zhu1, Colin Cherry "A Dataset for Detecting Stance in Tweets", 2016



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY