MINING OF SENTIMENT DATA ON INDIAN TOURISMS THROUGH TWITTER HANDLE

[1]Atufa Javeed, [2]P.Ramasubramanian

[1]PG Scholar, M.Tech, Dept of CSE, Shadan Women's College of Engineering and Technology HYD,T.S.
[2]Professor, Dept. of CSE, Shadan Women's College of Engineering and Technology, HYD, T.S.

ABSTRACT

Tourism plays most important role in the rise of countries for preserving cultural memory and approaching peace. Sentiment analysis helps in analyzing the point of view or the emotion expressed by the people for different organizations or business applications. The author applies the Sentiment analysis to the review given by people on twitter about Indian tourism. Twitter is an online forum where users are allowed to share and post their thoughts in the format of tweets with a big audience. Sentiment Analysis on twitter helps in analyzing whether the emotion is positive, neutral or negative. The intent of the author is to examine people's observation about the Indian tourism by using twitter dataset and do online reviews of all the users and put forward a new model for developing the tourism business and they're by increasing business gains by managing websites. In this paper, the author recommended domain specific ontology based on sentiment analysis method and created a Indian tourism based ontology - ConceptNet and the production of Indian tourism analysis is done by using sentiment analysis algorithm. The analysis of the tweets is done in 4 steps: (1) gathering the data i.e. the comments from twitter commented by customers (2) pre-processing the unnecessary symbols or catch-phrases (3) identifying the review as positive, negative or neutral (4) summarizing the results into graphs. The paper proposes novel technique to perform these tasks.

I. INTRODUCTION

Online reputation is one of the most precious assets for brands. A bad review on social media can be costly to a company if not handled properly and swiftly. Twitter is a major source of customer insight. In fact, people use it to express all sort of feelings, observations, beliefs and opinions about a variety of topics. By performing Twitter sentiment analysis, we can unleash the power of this data and use it as a valuable asset.

Let's say a company has just launched a new product feature and it notices a sharp increase in mentions on Twitter. However, receiving tons of mentions doesn't necessarily mean a good thing. Are customers tweeting more because they're expressing good things about this new feature? Or are customer's actually complaining about the feature having lots of bugs? Performing Twitter sentiment analysis can be an excellent way to understand the tone of those mentions and obtain real-time insights on how users are perceiving the new product. In this paper, the author is concern with person's assumption about Indian travel industry since India is the delightful nation, all are intrigued to visit India. Since twitter is the huge corpus of information, twitter message has taken for this exploration to know the individual's notion. At last we present semantic examination idea which worry about the semantic importance of the word. It would be characterized in to two kinds relevant semantic and theoretical semantic. Logical semantic arrangements with considering about neighboring word. Calculated semantic of the word relies upon outside information, for example, ontologies and semantic organization. Sentistrength is the openly accessible asset. It ascertains the positive and negative supposition score in short content. Positive notion esteem ranges from 1 (not positive) to 5 (incredibly certain) and negative assumption esteem ranges from - 1 (not negative) to - 5 (amazingly negative). It is basically created for web-based media web text. For instance, 'I appreciate playing in the recreation center however the dread of creepy crawlies prevents me from doing as such', the term appreciate gives positive score 3, the term dread gives negative score - 4. So, the general notion of the above sentence is max [3, - 4]. It yields - 1 which means negative assumption.

SentiWordNet is a notion dictionary holding extremity score of the supposition words. It has around 3 million words including things, action words, qualifiers, modifiers. For the most part terms in SentiWordNet arranged into four classes to be specific descriptor, qualifier, action word and thing. SentiWordNet was worked by utilizing WordNet, this depends on 'pack of Synet' model. This dictionary holds promoter word rundown and emoji query table. SentiWordNet has three scores.

The objective of this project is to show how sentimental analysis can help firms/organizations and business associations to get a handle on their client's criticism with respect to their item and improve the client experience.

II. RELATED WORK

Our work is closely related to Vallikannu Ramanathan and T.Meyyappan's work on semantic classification of reviews based on Oman tourism. They examine the effect of four factors that is domain specific ontology, entity specific opinion extraction, combined lexicon-based approach and conceptual semantic sentiment analysis to determine the sentiment analysis of tweets about Oman tourism. They experiment the analysis using Naive Bayes approach. They show that the accuracy achieved combining lexicon-based approach results into 79.43%. However, the performance using Naive Bayes is limited since Naive Bayes classifier makes a very strong assumption on the shape of data distribution, i.e. any two features are independent given the output class. Our work differs from theirs in this aspect i.e. we analyze the tweets using Semantic sentiment analysis algorithm.

Go et al [2] explored sentiment analysis with the development of Twitter as a wellspring of general conclusion. The creator utilized "emojis" for tweet marking, and utilized unigrams, bigrams and Part-of-Speech labels as highlights. Diverse grouping calculation, for example, Naive Bayes, SVM and Maximum Entropy (MaxEnt) have been utilized and analyzed. The creator tried their frameworks on physically and named test set of 359 tweets, and accomplished a maximum exactness of 83.0% utilizing MaxEnt classifier on unigram and bigram highlights. One significant test actually remains is treatment of close to nonpartisan tweets.

Garg and Chatterjee [13] used Naive Bayes and MaxEnt classifiers in a two-venture strategy to initially group target versus abstract (unbiased) tweets and, at that point in the subsequent advance to characterize target tweets into positive versus negative sentiment. The creators found that solitary advance classifiers outflank such strategies.

Boia et al [10] experimented on marking of tweets utilizing emojis. The creators demonstrated that the majority of the mistaken marking occurred for tweets containing impartial sentiment which were either given a positive or negative name. This shows that emoji-based marking recognizes positive and negative sentiment quite well, in any case, that isn't the situation with unbiased tweets.

Kiritchenko et al [8] describes the state-of-the-art sentiment analysis framework which recognizes the sentiment of short casual literary messages including tweets. Their framework utilized sentiment highlights got principally from tweet-explicit sentiment vocabularies. The creator additionally produced a different vocabulary for discredited words. Point-wise Mutual Information (PMI) was utilized as a measure to dole out the scores.

Our work is also related to but different from subjective genre classification, sentiment classification, text summarization and terminology finding. We discuss each of them below.

Current resources for sentiment analysis suffer many deficiencies:

- The problem in sentiment analyis is classifying the polarity of a given text at the sentence or feature/aspect level
- Whether the communicated assessment in a sentence or a substance include/viewpoint is positive, negative or neutral.
- Failed Naive Bayes technique - The principle impediment of Naive Bayes is the suspicion of autonomous indicator highlights. Naive Bayes certainly expects that all the characteristics are commonly autonomous. In actuality, it's practically inconceivable that we get a lot of indicators that are totally autonomous or each other.

III. PROPOSED METHODOLOGY

We propose novel method based on domain specific ontology. We constructed our own Indian tourism ontology.

1. Domain Specific Ontology

The engineering of ConceptNet is more appropriate to being steadily refreshed, being populated from various information sources, and looking in complex questions are important to find good judgment analogies.

In the proposed method, the utilization of ConceptNet to assemble space explicit philosophy for Indian Tourism, further creator augments the metaphysics by incorporating the ontologies of related areas for the best inclusion of explicit highlights.

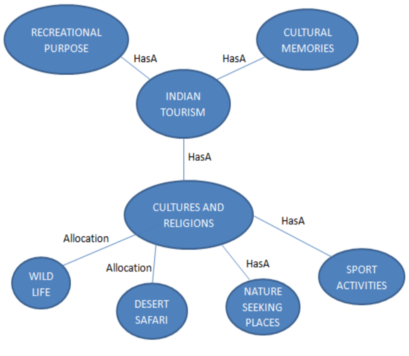


Fig 1: Domain specific ontology

In the above figure, the creator shows the formation of Ontology for Indian travel industry, the creator builds cosmology for 'Cultures and religions', at that point consolidate both the ontologies 'Indian the travel industry' and 'Social and religions by interfacing with new connection 'HasA'. Essentially, this can expand the travel industry philosophy by adding sport exercises and so on to make total Indian tour.

2.Entity Specific opinion extraction

Entities are recognized using the tourism ontology designed in domain specific ontology mentioned above. For instance, consider the tweet message 'India is the correct spot for wild safaris', wild safari is an element for which the sentiment is connoted. Ordinarily, thing is treated as an element. Here, wild safari is a thing and the word 'right' is the modifier which passes on the positive assessment. To follow out the substance in each tweet Part-of-speech labels are utilized.

Consider an example, 'Holy city of Varanasi'. Here, Varanasi and city are treated as elements since it watches the standard of all things to be considered as elements. After the extraction of things, these things are counted with area explicit cosmology – Indian the travel industry ideas. Superfluous highlights are eliminated with the assistance of cosmology. The assessment is finished by testing twitter with object-trait sets in metaphysics. So till now what we have done is ordered the words as certain or negative. How about we proceed onward ahead and ascertain the sentiment score, this is finished with the assistance of Lexicon based methodology.

3. Lexicon Based approach

This methodology helps in figuring the extremity score of the tweets. It depends on three existing sentiment vocabularies - SentiStrength is utilized for figuring the positive and negative sentiment score, SentiWordNet ascertains the extremity score of words and Opinion Lexicon-speaks to the most established sentiment word reference.

It ascertains the positive and negative sentiment score in short content. Positive sentiment esteem goes from 1 (not positive) to 5 (positive) and negative sentiment esteem goes from - 1 (not negative) to - 5 (negative).For instance, "the outing was extraordinary", so subsequent to cleaning and arranging the terms as certain and negative in this sentence, by the previously mentioned ideas, presently there are two words that is excursion and incredible. Here, incredible is positive word consequently give the worth 5 and the word trip is a nonpartisan word so 0 (for unbiased terms we dole out 0). So, the general sentiment of the above sentence is [1,0] which implies a positive sentiment.In view of instinct, the creator doles out the qualities of a couple of regularly utilized intensifiers and action words with values going from - 1 to +1. Consider the table below:

TABLE 1 Lexicon Based Approach

VERB	STRENGTH	ADVERB	STRENGTH
Love	1	complete	+1
adore	0.9	most	0.9
like	0.8	totally	0.8
enjoy	0.7	extremely	0.7
reject	-0.2	too	0.6
impress	0.5	any	-0.2
attract	0.4	not	-0.8
dislike	-0.7	less	-0.6

4. System Architecture

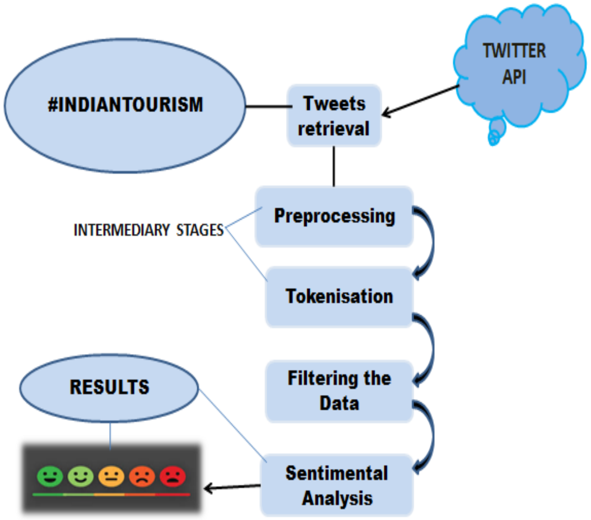


Fig 2: System Architecture

The proposed system consists of the following modules:

- Data gathering/ data acquisition
- Pre-processing
- Polarity Calculation
- Classification

♦ DATA ACQUISITION

The author assembles popular conclusion dependent on gathered hash label identified with sees about Indian the travel industry including Twitter top patterns, here it utilizes Tweepy API. The creator has made a record on Tweepy API connected to our Twitter account. To recover the tweets, Tweepy API acknowledges boundaries and gives the Twitter record's information consequently. Recovered tweets, from Twitter accounts, were spared in the information base under the accompanying fields: twitter_id, hashtag, tweet_created, user_id, screen_name, tweet_text, retweet_count, follower_count, and favourite_count of each tweet.

♦ PRE-PROCESSING

Twitter is a micro-blog where people generally Twitter is a miniature blog where individuals by and large write in a conversational style. Tweets are known to be loud for any content mining task as they contain various images that don't have any helpful data and make further processing ineffectual. Thusly this model incorporates successful pre-processing stage which eliminates unimportant images from tweets and henceforth, viable catchphrases can be separated. The means for pre-processing are as per the following:

i. Remove username and re-tweet symbol: Tweets often contain usernames starting with the image '@'. Once in a while a tweet is likewise re-tweeted, which implies a tweet by any client is shared again by different clients and it contains the image of retweet. These client names and re-tweet image don't contribute any criticalness to watchword extraction and go about as clamor. In this way, usernames and re-tweet images are eliminated.
ii. Remove URLs: Any URL joins showing up in the tweets are re-moved as the model zeros in just on the printed part of the tweet and URLs go about as superfluous commotion while catchphrases are extricated.
iii. Remove hash tags: The Hash tag i.e. # before a word such as #IndianTourismis removed to get 'Indian - Tourism'.
iv. Tokenization: Each term in a tweet is treated as a token. Tokens are the essential constituents of a tweet/text. Leave T alone the arrangement of tweets which is spoken to as T = {T1, T2, T3,…, T i | i is the number of tweets}. At that point each tweet in T is pre-handled and its terms are treated as tokens.
Let t be the set of tokens represented as t = {t1, t2, t3,…, tk }. t incorporates tokens from all the tweets of T where the quantity of tokens in the set T is k.
v. Stop word removal: A standard rundown of stop words is made and these stop words are then taken out from the set.

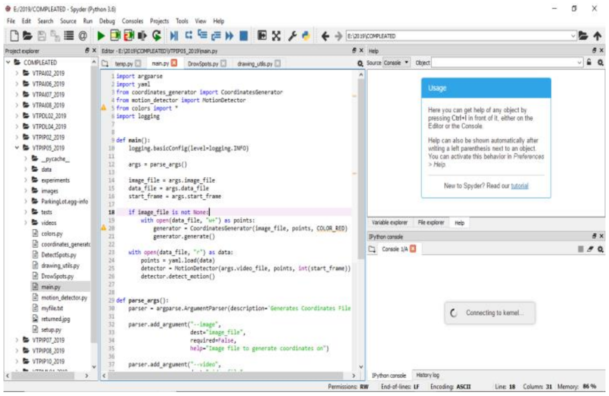
♦ POLARITY CALCULATION

Sentiment analysis can give important bits of knowledge from online media stages by recognizing feelings or assessments from an enormous volume of information present in unstructured configuration. Sentiment analysis incorporates three extremity classes, which are negative, nonpartisan and positive. The extremity of each tweet is controlled by relegating a score from −1 to 1 dependent on the words utilized, where a negative score implies a negative sentiment and a positive score implies a positive sentiment while the zero worth is viewed as an impartial sentiment. A score of subjectivity alloted to each tweet depends on whether it is speaking to an emotional importance or a goal meaning; the scope of subjectivity score is likewise from 0 to 1 where an incentive close to 0 speaks to objective and close to 1 abstract.

♦ CLASSIFICATION

Based on the polarity result we can classify the tweets into negative, positive and neutral reviews and also calculate the overall percentages. After getting the result we display the graphs based on the reviews.

IV. RESULT


E:\2019\COMPLEATED - Spyder (Python 3.6)
File Edit Search Source Run Debug Consoles Projects Tools View Help
Editor - E:\2019\COMPLEATED\VTPIP05_2019\main.py
temp.py
main.py
DrowSpots.py
drawing_utils.py
Project explorer
Help
Variable explorer
File explorer
IPython console
Console 1/A
History log

```
import argparse
import yaml
from coordinates_generator import CoordinatesGenerator
from motion_detector import MotionDetector
from colors import *
import logging


def main():
    logging.basicConfig(level=logging.INFO)

    args = parse_args()

    image_file = args.image_file
    data_file = args.data_file
    start_frame = args.start_frame

    if image_file is not None:
        with open(data_file, "w+") as points:
            generator = CoordinatesGenerator(image_file, points, COLOR_RED)
            generator.generate()

    with open(data_file, "r") as data:
        points = yaml.load(data)
        detector = MotionDetector(args.video_file, points, int(start_frame))
        detector.detect_motion()


def parse_args():
    parser = argparse.ArgumentParser(description='Generates Coordinates File

    parser.add_argument("--image",
                        dest="image_file",
                        required=False,
                        help="Image file to generate coordinates on")

    parser.add_argument("--video",
```

Usage

Here you can get help of any object by pressing Ctrl+I in front of it, either on the Editor or the Console.

Help can also be shown automatically after writing a left parenthesis next to an object. You can activate this behavior in Preferences > Help.

New to Spyder? Read our tutorial

Connecting to kernel...

Permissions: RW End-of-lines: LF Encoding: ASCII Line: 18 Column: 31 Memory: 86 %

Fig 3: Spyder Environment

```
           review
sentiment
negative      157
neutral       240
positive     1905
```

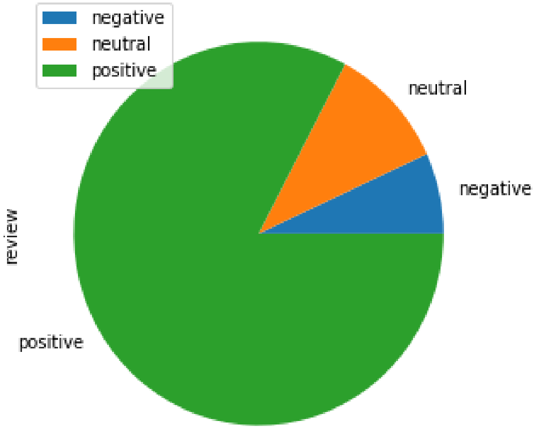


Fig 4: Sentiment Analysis

The figure above shows the pie chart that was used to justify the various percentages of tweets that bore positive, negative or neutral sentiments. It can be noted that the more than half of the tweets are positive about Indian tourism which represents that people have approved the tourism in India. The neutral sector denotes those tweets which either were not so clear with their stance on the Indian tourism or probably were indifferent about the tourism to form an opinion. The negative sector denotes the view that the people disapproved of the Indian tourism and expressed dismay at it. The experimental studies performed successfully show that the sentimental analysis and lexical analysis techniques for sentiment classification yield comparatively outperforming accurate results.

V. CONCLUSION

The aim of this report is focused mainly to develop a system that performs sentimental analysis on travel and tourism domain. In this project, the author highlighted the importance of the processing of the unstructured information provided in social networks inorder to determine the feeling that a user has towards the Indian tourism. In the future work, an adaptive ensemble member weighting process is designed to emphasize the importance of different ensemble members, and avoid the effect of deleterious ensemble members. NRC Word-Emotion Association Lexicon can be used for calculation of emotion of the tweet It is a collection of lexicons that was entirely created by experts of National Research Council of Canada. This lexicon collection can be used in a multitude of contexts such as sentiment analysis, product marketing, consumer behavior, and even political campaign analysis. The technology uses a list of words that help identify emotions, sentiment, as well as analyzing hashtags, emoticons and word-colour associations.

The future internet comes with high requirements of information dissemination, which motivate the research community to find alternative solutions.

REFERENCES

[1] A. Esuli and F. Sebastiani. 'SentiWordNet: a publicly available lexical resource for opinion mining' in Proceeding of the 5th International conference on Language Resources and Evaluation, pp.417-422, 2006.

[2] Alec Go, Richa Bhayani, Lei Huang "Twitter Sentiment Classification using distant supervision", CS224N Project Report, Stanford 12, January 2009

[3] Bing Liu, "Sentiment Analysis and Opinion Mining" Handbook of natural language Processing, 2010.

[4] Bing Liu, Minging Hu and Junsheng Cheng. 'Opinion Observer: Analyzing and Comparing Opinions on the Web' Proceedings of the 14th International World Wide Web conference (WWW-2005), May 10-14, 2005, Chiba, Japan.

[5] C. Havasi, R. Speer, and J. Alonso, 'Concept net 3: a flexible, multilingual semantic network for common sense knowledge' in recent Advances in Natural Language Processing, pp.27-29,2007.

[6] Erik Cambria 'An Introduction to concept-level sentiment analysis', In Advances in Soft Computing and Its Applications, pages 478-483. Springer, 2013.

[7] John R. Firth. A Synopsis of linguistic theory. Studies in Linguistic Analysis, 1930-1955.

[8] Kiritchenko, S.Zhu, X., & Mohammad S.M (2014), "Sentiment analysis of short Informal text", Journal of artificial Intelligence Research, 723-782, 2014

[9] Minging Hu and Bing Liu. 'Mining and Summarizing Customer Reviews', Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA.

[10] Marina Boia, Claudiu Cristian Musat and BoiFaltings, "How people attach sentiment to emoticons and words in tweets", Proceeding of the 2013 IEEE international conference on Social Computing, pages: 345- 350, 2013

[11] P. Turney 'Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL'02), Philadelphia, Pennsylvania, 2002.

[12] Vallikannu Ramanathan, T.Meyyappan 'An Exhaustive Exploration on Twitter Sentiment Analysis', Journal of Computer Science and Applications, International Research Publication house, Volume 6, Number 1, 2014. ISSN: 2231 -1270

[13] Yogesh Garg, Niladri Chatterjee, "Sentiment Analysis of Twitter Feeds in Big Data analytics", Springer International Publishing, pp 33-52 Dec 2014

AUTHOR'S PROFILE

Ms. ATUFA JAVEED has completed her B.Tech (CSE) from Kakatiya Institute of Science and Technology for Women, JNTUH, Nizamabad, T.S. Presently, she is pursuing her Masters in Computer Science & Engineering from Shadan Women's College of Engineering & Technology, Khairtabad, Hyderabad, T.S, India.

Prof. Dr. P. RAMASUBRAMANIAN is presently serving as Professor in the department of Computer Science and Engineering in Shadan Women's College of Engineering and Technology, Khairtabad, Hyderabad, India. He obtained his bachelor and master degree in Computer Science and Engineering and Ph.D. degree in Computer Science from Madurai Kamaraj University, Madurai in the year 1989, 1996 and 2012 respectively. He has more than 30 years of teaching experience, authored 16 books, published 47 research papers in international, national journals & conferences and 111 citations with h-index 5 and i10-index 4 in his credit. He has guided two Ph.D research scholars under Bharathiar University and degree was also awarded. His current area of research includes Data Mining, Data Ware housing, Neural Networks, Fuzzy, Rough Set logic and Emotional Intelligence. He is a member of various professional societies like Indian Society for Technical Education, International Association of Engineers, Computer Science Teachers Association, International association of Computer Science and Information Technology, Fellow in Institution of Engineers (India), and Fellow in international Society of Research and development. Prof. Dr. P. Ramasubramanian has been chosen who's who in Science and Engineering in the year 2011. He is a reviewer and editor for various international journals and conferences.