Crowd Prediction System for Tourists

Kartik Rajendra Kokane
dept. Computer Engineering
APSIT
Thane, India
kartikk981@gmail.com

Pratik Pradeep Jogdand
dept. Computer Engineering
APSIT
Thane, India
jogdand343@gmail.com

Rohit Dhuri
dept. Computer Engineering
APSIT
Thane, India
rohit.dhr43@gmail.com

***Abstract*—Tourism Industry is one of the largest industries in the world with a global economic contribution of over 7.6 trillion USD in 2016 as mentioned on statista website [1]. One of the major problems faced by tourist all over the world right now is that of crowd. Popular tourist destinations tend to get extremely crowded and tourists plan their trips without any insights. Research solutions till date propose results based on static data from previous years which is not cost effective. This model plans to provide a report predicting the crowd density of particular location with respect to the dates mentioned by the user. After taking a glance at the report the user can decide for himself, whether to proceed with the plan or pick up a new location. The model can also suggest similar locations as per the expectations of the user which may be less crowded during the time frame provided.**

***Index Terms*—Crowd density, sarcasm analysis, hashtag processing, sentiment analysis, Application programming interface, Absolute Percentage Error, Recommender Systems.**

I. Introduction

All over the globe, there has been a steady growth in the tourism industry where International tourist arrivals increased from 528 million USD in 2005 to a staggering 1.19 billion USD in 2015 [1]. In the past couple of decades, we witnessed an accelerated technological progress characterized not only by new innovations, but their application and diffusion also mattered.

The word 'Tourist' is coined for individuals traveling for recreation. Growing amount of tourists at famous locations can now be seen as an obstacle for recreation. With new technologies which are available right now, it is possible to predict the crowd density of any given location with respect to the dates provided. The average crowd density for a frame of 3 months will also be calculated with the dates provided serving as the median for the frame. Then both the values i.e average value and the predicted value will be compared for the report to be generated. Deviation of the predicted value and the average value will help us in predicting crowd density of the given time frame.

Crowd occurs due to regional exquisites, festivals and important events which attracts large crowd that demands immediate and instantaneous service. Understanding the crowd is a difficult task, therefore our data set will use information from Social Media, Hotels and Travel companies. Most of the websites nowadays share some of their information for public use which can be used through APIs (Application programming interface). Hashtags, location updates and mentions from social media platforms such as Facebook, Twitter and Instagram will considered as input sources. Availability and prices of respective rooms of hotels at the selected location. Flight, train fares and seat availability will also be considered. Averaging out all the listed data set values and then comparing with output values of the desired date, the model will try to predict the crowd density.

One of the major applications of machine learning in business is Recommender Systems (RS). Recommender systems are nothing but a big filter which searches through large volume of dynamically generated data and provides users with content which is personalized exclusively for them. [9][10][11]. Therefore, a subclass of our system is to recommend similar location if the average crowd density is excessive as per the requirements of the user.

A. Aim

A crowd density predicting model which will understand information provided by the datasets and then predict the crowd footfall at the requested location on provided dates by the user. If it gets too crowded then the model will suggest similar locations which will have less crowd but same characteristics.

B. Objective

The primary objective is to create an automated system that will fetch information from different platforms such as social media, travel websites and then understand it to predict the crowd footfall at requested location on provided dates by the user. To attain this successfully, the system must meet the following goals.

1) To retrieve data dynamically from various platforms mentioned above to form the datasets for the system.
2) To understand the datasets and then calculate the **crowd density** at the requested location on provided dates by the user.
3) To provide multiple alternative locations if the expected crowd footfall is much more than the average footfall.

II. Design

We propose to implement this system on two use cases namely, hotel and travel bookings and social networking sites(Twitter and Instagram). And choose the use case which yields the most accurate**crowd density**.

A. Hotel and travel bookings.

During times of festivals or regional specialties there is an increase in the population density. This is directly proportional to hotel fares and travel ticket fares. Our system analyses this difference between the yearly average price and the increased price. Based on the result obtained, the system predicts probability of crowd density at the given time.

B. Social network

Social networking sites have spread like a virus and have become an important aspect of our society. Users share every moment of their lives on these platforms and our system aims to use this data to predict their activity. The system learns from dynamic datasets obtained from Twitter and Instagram and helps in determining which content is relevant to the users search. To classify this content into different categories we compare it with various choices of evaluation such as hashtag processing, sentiment analysis and sarcasm analysis.

C. Sentiment Analysis

Systems Sentiment Analyzer [2][3] classifies tweets into three categories positive, negative and neutral. The goal of this analysis is to determine whether the positive sentiment has a different effect on the prediction result than negative sentiments.

D. Hashtag processing

Earlier, hashtags have been used as approximate topic indicators for tweets and instagram posts. [4][5][6][7]. We will use the volume of certain hashtags and discrete relevant topics from the entire pool.

E. Sarcasm Analysis

The model will use Part of speech(POS) [3][8] tags that will check the credibility by comparing tweet sentiment with its context. The system will assign sarcastic and non-sarcastic tag for each tweet and dump the sarcastic tweets.

III. Working

Upon landing on the interface the users will be asked to enter location and the time frame of their travel. The system will receive data from the users input. Then the system will start fetching data from different platforms which are mentioned in this paper to generate information as a source for the data set. Once all the values are embedded in the dataset, the systems algorithm will analyze the dataset using Machine learning technologies to calculate a possible value or the volume of the particular values at that location for the given time frame. On the other hand the system will also figure out the annual average values simultaneously. The newly calculated value will be compared with the annual average values and the deviation of these values will help us determine the crowd density. If the insights predict that the crowd density is much more than the average values, then the system will suggest alternate locations having similar characteristics which may have less crowd.

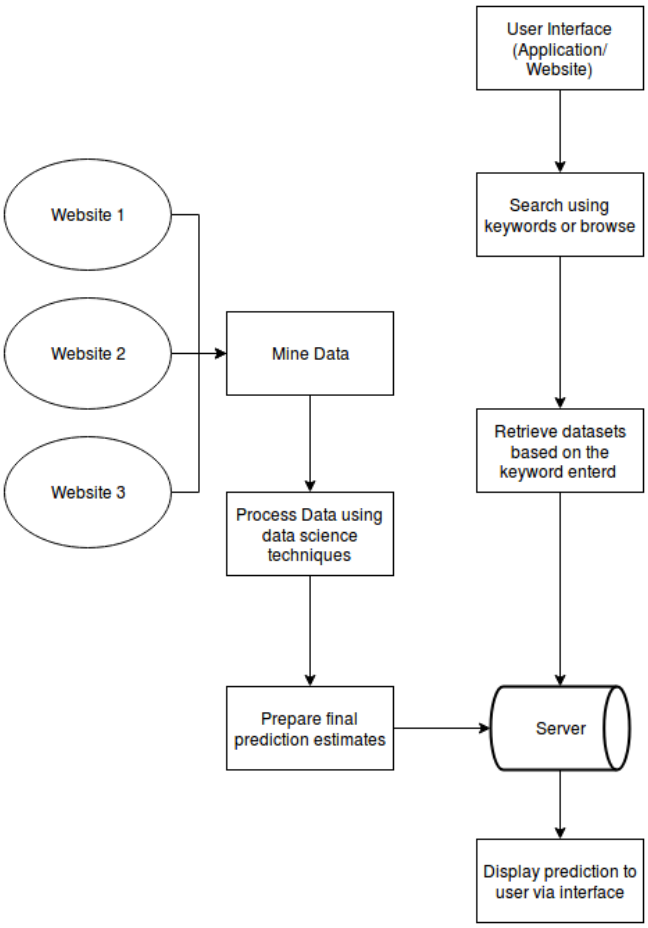

Fig. 1. Estimated Methodology.

For example, if the annual average values of a particular location is 50 units. And the newly calculated value is 80 units. Then the user may expect 160% crowd attending the location on chosen dates where 100% being the average crowd.

IV. Literature Review

To generate personalized insights for the user, RS is important as it can help tourists save energy and time while planning for their trip.[12][13]. Now coming to social media, Twitter is a famous micro-blogging platform. On twitter people share tweets of maximum 140 characters. Generalizing all the tweets, we can further divide them broadly into two categories. First being the tweets where the user tweets about themselves and second where information is shared using the tweet[14] This information shared is a very important source for collecting data and determining twitter sentiments.

Sentiment analysis is useful in commercial intelligence application environment and recommender systems [19], [20]. B. Liu (2012) defined sentiment analysis, which is also called opinion mining as the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organi-

zations, individuals, issues, events, topics, and their attributes [18]. Based on the definition, data can be then sub-divided into positive, negative or neutral.

Liebrecht et al. [15] discusses how sarcasm plays an important role of changing the polarity of a message and how it conveys a negative opinion using only positive words. Therefore, detection of sarcasm is important for the development of sentiment analysis systems.

Maynard and Greenwood [16] have put forward a set of rules to determine whether a tweet is positive or negative when sarcasm is detected. Hashtags are used as labels on the internet. Hashtags gained popularity on twitter itself [17]. Maynard and Greenwood [16] discuss how hashtags contain much useful sentiment information, but how it is difficult to identify tags as they contain multiple words. They have developed an algorithm to extract the individual words from a single hashtag.

V. Conclusion

A new approach for predicting crowd density is proposed by using this prototype model, which could be used as an application to grow tourism by suggesting the appropriate locations for the right time frame. It uses dynamic values, rather than static values of data as source of information for generating datasets. Implementing machine learning technologies on these data sets, the model will be able to predict the crowd density. A subordinate function will execute if the crowd density surpasses the annual average values providing with an alternative location with same characteristics. Depending on the success rate of use cases, we propose two strategies, namely social media and travel fare prices.

Future Scope : For future work, we are going to develop algorithms to process datasets and predict crowd density using machine learning. Keeping availability of rooms and travel options as a determining factor, we will also research on various other factors such as Hashtag processing, Sentiment analysis, Sarcasm analysis, Location updates, mentions on Social media and how they will help us attain maximum accuracy in predicting crowd density.

References

[1] Statista- The Statistics Portal https://www.statista.com/topics/962/global-tourism/.

[2] Hassan Saif et al., "Semantic patterns for sentiment analysis of Twitter", *in International Semantic Web Conference, Springer International Publishing* 2014.

[3] Nitesh Kumar Singh, Om Kumar C.U, Dr. Rajeshwari Shridhar., "Flash Crowd Prediction in Twitter.", *International Conference on Advanced Computing and Communications Systems* 2017.

[4] Dagmar Gromann, Thierry Declerck., "Hashtag PRocessing for Enhanced Clustering of Tweets"

[5] Kevin Dela Rosa, Rushin Shah, Bo Lin, Anatole Gershman, and Robert Frederking. , "Topical clustering of tweets", *Proceedings of the ACM SIGIR:SWSM* 2011.

[6] Pavan Kapanipathi, Prateek Jain, Chitra Venkatara-mani, and Amit Sheth., "User interests identification on twitter using a hierarchical knowledge base", *In European Semantic Web Conference.Springer, pages 99113* 2014.

[7] Piyush Bansal, Romil Bansal, and Vasudeva Varma. 2015, "Towards deep semantic analysis of hashtags", *In European Conference on Information Retrieval. Springer, pages 453464.* 2015.

[8] S. K. Bharti et al., "Towards deep semantic analysis of hashtags", *In European Conference on Information Retrieval. Springer, pages 453464.* 2015.

[9] Mahmood, T., Ricci, F, "Improving recommender systems with adaptive conversational strategies", *In: C. Cattuto, G. Ruffo, F. Menczer (eds.) Hypertext, pp. 73 82. ACM* 2009.

[10] Resnick, P., Varian, H.R, " H.R.: Recommender systems", *Communications of the ACM 40(3), 5658.* 1997.

[11] Burke, R., "Hybrid web recommender systems", *In The AdaptiveWeb, pp. 377408. Springer Berlin / Heidelberg* 2007.

[12] A.Moreno, L.Sebasti and P.Vansteenwegen, "Recommender Systems in Tourism", 2015.

[13] Kevin Meehan, Tom Lunney, Kevin Curran, Aiden McCaughey, "Context-Aware Intelligent Recommendation System for Tourism", 2013.

[14] Johan Bollen, Huina Mao, and Alberto Pepe., "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena", *In International AAAI Conference on Weblogs and Social Media (ICWSM'11)* 2011.

[15] C. C. Liebrecht, F. A. Kunneman, and A. P. J. van den Bosh, "The perfect solution for detecting sarcasm in tweets *hashtag*not ", *In Proc. WASSA, pp. 2937* , Jun. 2013.

[16] D. Maynard and M. A. Greenwood, "Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis", *In Proc. 9th Int. Conf. Lang. Resour. Eval., pp. 42384243.* May 2014.

[17] Godin, Frderic, Viktor Slavkovikj, Wesley De Neve, Benjamin Schrauwen, and Rik Van de Walle., "Using topic models for twitter hashtag recommendation", *In Proceedings of the 22nd International Conference on World Wide Web, pp. 593-596.* ACM, 2013.

[18] Bing Liu, "Sentiment Analysis and Opinion Mining", *In Morgan and Claypool Publishers* May 2012.

[19] L. Terveen, W. Hill, B. Amento, D. McDonald, and J. Creter, "PHOAKS: A system for sharing recommendations", *In Commun. ACM, vol. 40, no. 3, pp. 5962*, 1997.

[20] J. Tatemura, "Virtual reviewers for collaborative exploration of movie reviews", *In Proc. 5th Int. Conf. Intell. User Interfaces, 2000, pp. 272275.* 2000.