Opinion Ensembling to Analyse Economic Growth through Tourism

Dr. C. S. Kanimozhiselvi[1] P. Keerthana[2] A. KishoreKumar[3] D. Kowtham[4]
[1,2,3,4]Department of Computer Science and Engineering
[1,2,3,4]Kongu Engineering College, Perundurai, India

Abstract— Throughout the world, tourism brings money to cities and countries. Tourism also provides jobs for the local residents, further benefiting the destination. India has realized the profits available from this sector.Online reviews had a noticeable impact over the tourism industry, as tourists now make decisions based upon the reviews written by the tourists on travel websites. Online reviews directly or indirectly affects the economy of the country through foreign investment or through job opportunities. To process these online reviews as opinions faster, many algorithms were designed which changed the way people viewed online reviews. However, the individual results of these opinion processing algorithms were found biased or varying in some cases, which gave rise to the need of opinion ensemble techniques. In this paper three different tourism types are taken and processed the opinions of each type and analyzed the economic growth. We used opining ensembling technique to evaluate the results. The reviews are extracted from tripadvisor.in. The empirical evaluation of the extracted reviews has resulted in suggestions to improve economic growth through tourism.

Key words: POS Tagging, Sentiment analysis, Opinion mining, Tourism, Ensembling etc

I. INTRODUCTION

Opinions have been an important part of our lives and it is a human behavior to analyze these opinions to take decisions. The internet has emerged as a vast retainer of travel data, where users add latest travel reviews every day. TripAdvisor is one of the largest and most visited travel & tourism websites, with the database of more than 60 million members and over 170 million reviews and opinions for hotels, restaurants, attractions and other travel-related businesses [1]. Such online reviews form a sizable part of it and have a variety of information stored in them. Travellers go through these reviews and seek the necessary information they need. Such huge load of data makes it impossible for single person to read it all which gave rise to need for collecting and processing those reviews, and summarizing users relevant information. Several websites are equipped with rating system that grades reviews or other data by stars/numbers or text, while some websites cater both text as well as rating system [2]. Using only numerical rating system does not grant enough data as there are many other review related problems, which makes it hard to assess. Some of them are:

- Lengthy reviews about the places makes the reader to ignore the review,
- Opinions differ from one user to another
- Overall rating is affected by multiple aspects

It is the field that deals in determination and classification of opinions or feelings expressed in review [3]. Several algorithms have been designed to process these online reviews as opinions. However, the individual results of these opinion processing algorithms were found biased or varying in some cases, which gave rise to the need of opinion ensemble techniques. Ensemble is an approach that epitomizes the review and excerpt the opinions from the data which gives the main context. Ensemble methods train multiple learners on the provided data set to solve the same problem then combine them to form a single model [3]. Thus, opinions mined through ensemble approach extricate the travellers from decision making process. Henceforth, impacting the economic growth as tourism provides direct and indirect jobs to people.

The remainder sections of the paper are organized as follows: previous work is expounded in Section 2; tourism and economic growth is discussed in section 3; the methodology explained in Section 4; an experimental evaluation of opinion ensemble approaches in the online review is performed in Section 5; and finally, conclusions and future work are discussed in Section 6.

II. PREVIOUS WORK

In [4] Khan and et al. presented literature survey of opinion mining. They have summarized various ma-chine learning algorithms for sentiment classification from unstructured reviews. They have discussed various applications of opinion mining such as search engines, recommendation systems, email filtering, Web ad filter-ing, questioning/answering systems. In [5] Cristian Bucurab proposed a system for extracting and summarizing opinions expressed by users on TripAdvisor.com. Sentences are separated into integral units as words using the tokenization process and SentiWordNet evaluates the polarity of the separated words. They have evaluated the platform using text mining domain specific measures such as recall, f-measure, accuracy and precision. In [6] Cambria and et al. have created a compilation of frequently used polarity concepts i.e. common approach with relatively strong positive or negative polarity. They have developed SenticNet, a publicly available semantic re-source for opinion mining. It exploits common argumentative techniques, such as blending and spectral activation, together with an emotion categorization model and an ontology for describing human emotions. In [7] Hu and et al. proposed a multi-text summarization technique for identifying the top- k most informative sentences of hotel reviews posted on TripAdvisor.com. To determine the similarity of two sentences the content and sentiment similarities were used. The k-medoids clustering algorithm was used to partition sentences into k groups and iden-tified the top-k sentences. The medoids from these groups were then selected as the final summarization results. In [8] Taylor and et al. have extended Bing Liu's approach to describe customer inclination regarding tourism products by making use of the aspect-based opinion mining approach. They have extracted product reviews from Los Lagos, particularly, hotels and restaurants. They measured the performance of the proposed algorithm and designed and developed an application to extract opinions from these reviews and to generate proposed sum-marization charts. In [9] Lin and Chao proposed an approach

for tourism-related opinion detection and tourist attraction target identification. They have extracted blog articles labeled as in the domestic tourism category in a blogspace. Annotators were used to annotate the opinion polarity and the opinion target for every sentence. They have used machine learning methods to train learners for tourism-related opinion mining. In [10] Basari and et al. have proposed a hybrid method of support vector machine and particle swarm optimization for opinion mining of movie review. A SVM-PSO technique improved the parameters of SVM using PSO. In [11] Li and et al. have developed VisTravel: visualizing tourism network opinion from the user generated content. They have extracted e-tourism User-Generated Content data from Mafengwo, which is one of the Chinese travel social networks. In [12] Akehurst have discussed the importance to systematically identify the type of tourist or traveller who actually writes blogs and what types of trip and stays in destinations are more likely to generate meaningful User Generat-ed Content. In [13] Godnov and Redek have discussed the case study of Croatia for text mining in tourism. Latent Dirichlet Allocation was used to identify topics discussed in texts. It identified the nodes of each word cluster and accompanying words. In [14] Taylor and et al. have proposed OpinionZoom, a modular software. It helps users in an effective way to understand the vast amount of tourism opinions disposed all over theWeb. They have tested OpinionZoom, encompassing the situation of the tourism industry in Los Lagos, also known as the Lake District, in Chile.

III. ECONOMIC IMPACT OF TOURISM

In the recent time, the tourism has entered a new era where almost all activities are controlled via the world-wide-web. Information technology has made its impact upon tourism also and the industry has welcomed and adopted it with equal respect. With the rise of internet, another platform, known as social media, has also made an appealing and strong presence alongside tourism reviewing websites. Tourism industry is using the social media platform for destination marketing by setting up accounts on social networks and using them to attract customers[15]. Tourism industry hires special team that specializes in this type of marketing since it has shown positive results. This acts as an effective and easy way of promoting a business which targets users based upon their interest instead of random advertisements [16].

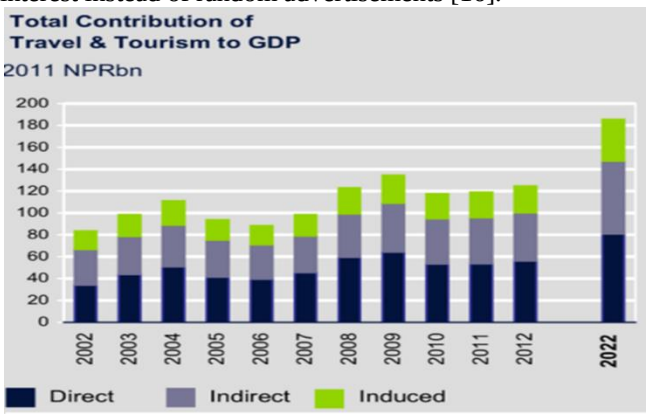

Fig. 1: Economic effect of Tourism in terms of GDP

Tourism is among the worlds largest and fastest expanding industries with significant impact on economy. For various countries tourism is the main source of promoting local development by provocative economic activities and statistics analysis of tourism economy gives an estimate of growth in the country[17]. Tourism has its direct economic impact, in terms of GDP generated by services that deal directly with tourists. Hotels, restaurants, travel agents, airlines, other passenger and cargo transport and leisure industries impact economy in terms of GDP. Figure 1 depicts the impact of tourism on economic growth for Croatia. Flora and Fauna are the major attractions for tourists, for tourism based on the natural habitat and historical and cultural heritage. Tourism industry provides the youth with employment opportunities and earns a hefty revenue as well.

IV. METHODOLOGY

This section is comprised of the dataset description, the preprocessing procedure, and the classification algorithm. All the experimental processes have been completed using the RSTUDIO.

A. Dataset Description

This paper works with three different types of tourism datasets, each consisting of reviews and the labels. The detailed description is shown in table 1.

S.NO	TYPES OF DATASETS	TOTAL REVIEWS
1	Pilgrimages	2500
2	Adventure Tourism	2350
3	Heritage Tourism	2435

Table 1: Datasets used

B. Data Preprocessing

Data undergoes following preprocessing steps such as

- Tokenization: Tokenization is a step which splits longer strings of text into smaller pieces, or tokens. Larger chunks of text can be tokenized into sentences, sentences can be tokenized into words, etc.
- Stemming: It is the process of removing suffixes from words to get the common origin. In statistical analysis, it greatly helps when comparing texts to be able to identify words with a common meaning and form as being identical.
- Stopword removal: Stop words are words which are filtered out before or after processing of natural language data (text). ... Some tools specifically avoid removing these stop words to support phrase search.
- POS Tagging: It is the process of marking up a word in a text (corpus) as corresponding to a particular part of speech, based on both its definition and its context.

C. Wordcloud Formation

A word cloud is a graphical representation of frequently used words in a collection of text files. The height of each word in this picture is an indication of frequency of occurrence of the word in the entire text. Such diagrams are very useful when doing text analytics. Fig 3 shows word cloud formation.

Fig. 3: Word cloud formation

D. Opinion Ensembling

Most previous studies on opinion mining of reviews have employed a single learner on text review. The predictions given by an individual learner are often biased on a data set or have high variance factor due to their mode of implementation. Ensemble is advantageous as it uses different set of learners which makes it unlikely for all learners to misconstrue. If a few learners muff the calculations but the error is made only by some learners, the result will still be optimal as the inductive biases of different learning algorithms are highly correlated which imply that various algorithms are sensitive to identical errors. Ensemble methods contribute in reducing this biased variance of learners [18]. Hence, ensemble method are useful in cases where the learner algorithms are conscious of minor alterations in the training data set.

Ensemble Algorithms	Reference	Description
Logistic regression	Cox, DR (1958). "The regression analysis of binary sequences (with discussion)". J Roy Stat Soc B. 20: 215242. JSTOR 2983890.	A regression model where the dependent variable is categorical.
Random Forests	L. Breiman, Random Forests, Machine Learning, 45(1), 5-32, 2001.	Random Forests are an improvement over bagged decision trees.
Gauss Naive Bayes	Mitchell, T (1997). Machine Learning, McGraw Hill.	Can perform online updates to model parameters via partial fit method
SVC	Wang H., Yang Y.,Wang H., Chen D. (2013) Soft-Voting Clustering En-semble. In: Zhou ZH., Roli F., Kittler J. (eds) Multiple learner Systems. MCS 2013. Lecture Notes in Computer Science, vol 7872. Springer, Berlin, Heidelberg	Soft Voting/Majority Rule learner for unfitted estimators.
Bern Naive Bayes	C.D. Manning, P. Raghavan and H. Schuetze (2008). Introduction to Information Retrieval. Cambridge University Press, pp. 234-265.	Like MultinomialNB, this learner is suitable for discrete data. The d-ifference is that while MultinomialNB works with occurrence counts, BernoulliNB is designed for binary/boolean features.
Bagging	L. Breiman, Bagging predictors, Machine Learning, 24(2), 123-140, 1996.	A Bagging learner is an ensemble meta-estimator that fits base learners each on random subsets of the original dataset and then aggregate their individual predictions (either by voting or by averaging) to form a final prediction.
Extra Trees	P. Geurts, D. Ernst., and L. Wehenkel, Extremely randomized trees, Machine Learning, 63(1), 3-42, 2006.	This class implements a meta estimator that fits a number of randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting.
AdaBoost	Y. Freund, R. Schapire, A Decision-Theoretic Generalization of on- Line Learning and an Application to Boosting, 1995.	An AdaBoost [1] learner is a meta-estimator that begins by fitting a learner on the original dataset and then fits additional copies of the learner on the same dataset but where the weights of incorrectly classi-fied instances are adjusted such that subsequent learners focus more on difficult cases.
GradientBoostinglearner	J. Friedman, Stochastic Gradient Boosting, 1999	GB builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions.

Table 2: Ensembling Algorithms

Ensembling works in two steps (1) training the base learners, and then combining them. Figure 2 describes the the proposed opinion ensemble.

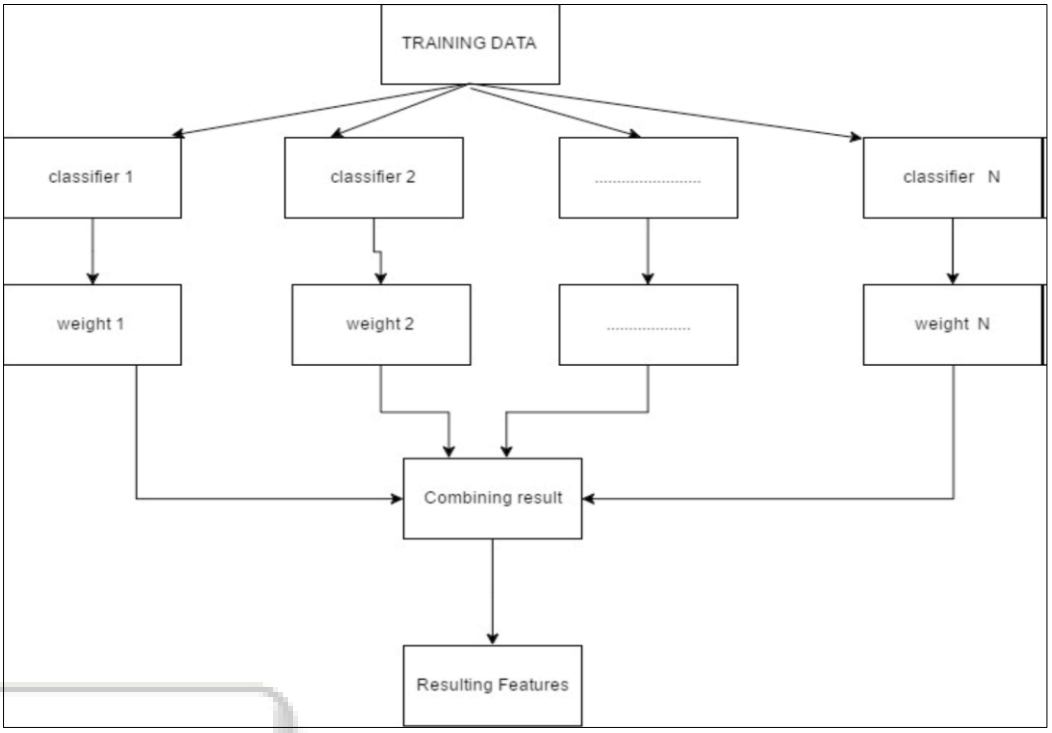

Fig. 2: Opinion Ensemble

Opinion Ensemble train multiple learners to solve the same problem. In contrast to other learning approach-es which try to construct one learner from training data, opinion ensemble try to construct a set of learners and combine them. Opinion ensemble contains several learners called base learners. Base learners are generated from training data by a base learning algorithm which can be decision tree, random forest or other kinds of learning algorithms. Table 1 summarizes the popular learners such as Logistic regression, Gauss Naive Bayes, Random Forests, Decision Tree, SVC, Bern Naive Bayes, Bagging, Extra Trees, AdaBoost, and GradientBoostinglearner which can be used as base learners. Opinion ensemble use a single base learning algorithm to produce homo-geneous base learners, i.e., learners of the same type, leading to homogeneous ensembles. The review training data is divided into N different data sets. It trains each learner on the review train data set. The weightage is assigned to result of each learner and finally by voting their results are ensembled. Opinion ensemble generalizes the individual learners and works better than that of base learners as it is able to boost weak learners.

V. Experiment Analysis

We used both the accuracy and sentiment classification results to measures to compare opinions among the tourist places. We used Random Forests, Decision Tree, SVM, Bagging algorithms to improve the classification accuracy. Metrics considered are Accuracy, Sensitivity and sentiment classification results.

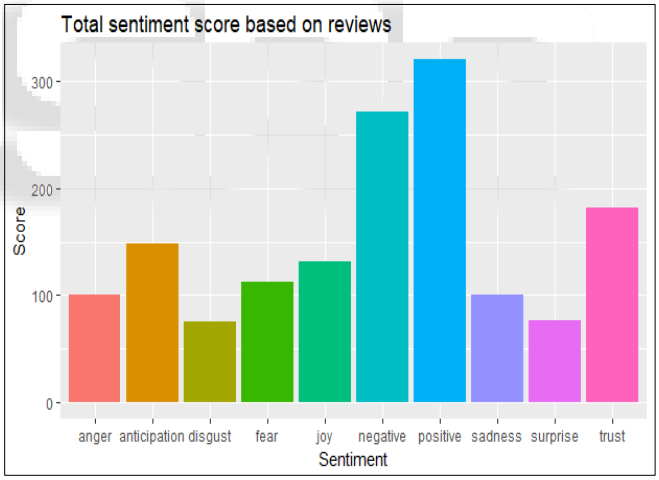

Fig. 3: Pilgrimage sentiment result

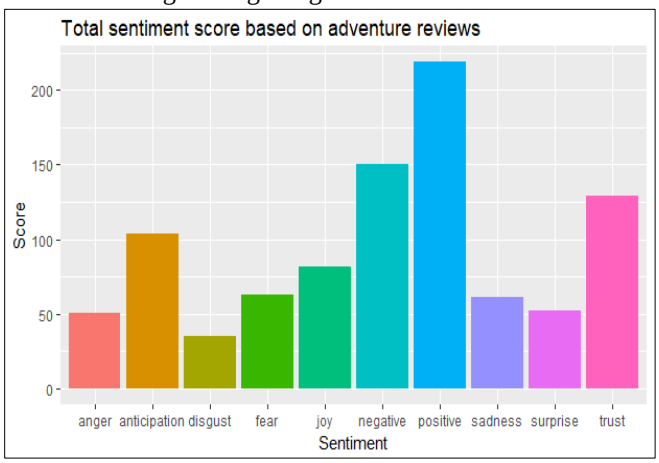

Fig. 4: Adventure sentiment result

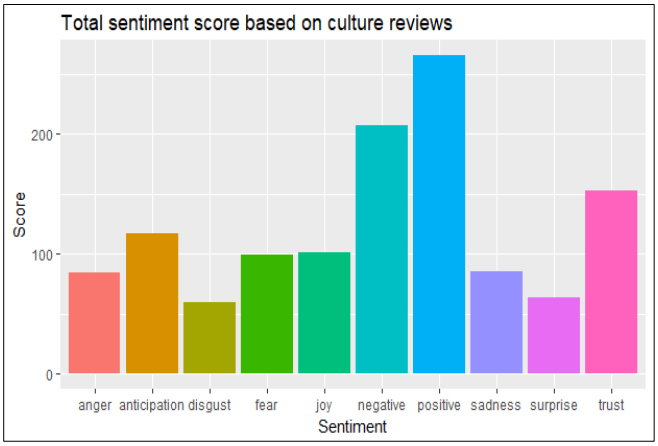

Fig. 5: Culture sentiment result

VI. CONCLUSION

We propose opinion ensembling to determine the opinions impacting economic growth from online tourism reviews. As analysing the sentiment results of three tourism Pilgrimages provides more positive polarity than other two. So this contributes more for economic growth. In opinion ensemble, multiple learners are trained and combined to classify online reviews. Furthermore, we extracted the online reviews from tripadvisor.com using scrappy. The results indicated that the important opinions can provide comprehensive information on economic growth due to tourism. Future research can utilize the reviews so as the study the different aspects of economic growth on tourism.

REFERENCE

[1] B. Garth, Global Flies into the Hall of Fame, www.globalballooning.com.
[2] T. R. UroÅ Godnov, Application of text mining in tourism: Case of Croatia.
[3] Kanimozhiselvi CS, Tamilarasi A. Mining of High Confidence Rare Association Rules with Automated Support Thresholds. European Journal of Scientific Research. 2011;52(2).
[4] A. K. A. U. Khairullah Khan, Baharum Baharudin, Mining opinion components from unstructured reviews: A review, Journal of King Saud University Computer and Information Sciences (2014) 26, 258275.
[5] C. Bucur, Using opinion mining techniques in tourism, Global Conference on Business, Economics, Management and Tourism.
[6] C. H. A. H. Erik Cambria, Robert Speer, Senticnet: A publicly available semantic resource for opinion mining, Commonsense Knowl-edge: Papers from the AAAI Fall Symposium (FS-10-02).
[7] H.-L. C. Ya-Han Hu, Yen-Liang Chen, Opinion mining from online hotel reviews a text summarization approach, Information Processing and Management.
[8] F. B.-M. Y. M. Edison Marrese-Taylor, Juan D. Velasquez, Identifying customer preferences about tourism products using an aspect-based opinion mining approach, Procedia Computer Science 22 (2013) 182 191.
[9] C.-J. L. P.-H. Chao, Tourism-related opinion mining.
[10] I. G. P. A. J. Z. Abd. Samad Hasan Basari, Burairah Hussin, Opinion mining of movie review using hybrid method of support vector machine and particle swarm optimization, Procedia Engineering 53 (2013) 453 462.
[11] S. W. M. L. X. F. H. W. Qiusheng Li, Yadong Wu, Vistravel: visualizing tourism network opinion from the user generated content, The Visualization Society of Japan 2016.
[12] G. Akehurst, User generated content: the use of blogs for tourism organisations and tourism consumers.
[13] T. R. Uros Godnov, Application of text mining in tourism: Case of croatia, Annals of Tourism Research.
[14] F. B.-M. Edison Marrese-Taylor, Juan D. Velasquez, Opinionzoom, a modular tool to explore tourism opinions on the web, 2013 IEEE/WIC/ACM International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT).
[15] G. Akehurst, User generated content: the use of blogs for tourism organisations and tourism consumers.
[16] P. Madasu, Social Media Marketing and Promotion of Tourism, Manag. Insight ix (1) (2013) 71–80.
[17] S. a. Z. YingYu, YiruiWang, Statistical modeling and prediction for tourism economy using dendritic neural network.
[18] H. D. III, A course in machine learning.
[19] http://www.sltda.gov.lk/index.html.

