AUTOMATIC SENTIMENT SCORE GENERATION METHOD FOR SIGHTSPOTS REVIEW SYSTEM

Victor SILAA[1], Fumito MASUI[2] and Michal PTASZYNSKI[3]

[1]Doctoral Course in Manufacturing Engineering, Kitami Inst. of Tech.
E-mail: alexvictortz@gmail.com
[2]Professor, Dept. of Computer Science, Kitami Inst. of Tech.
E-mail: f-masui@mail.kitami-it.ac.jp
[3]Associate Professor, Dept. of Computer Science, Kitami Inst. of Tech. E-mail: E-
E-mail: michal@mail.kitami-it.ac.jp
(Koencho 165, Kitami, Hokkaido 090-8507, Japan)

When planning a travel or an adventure, sightseers increasingly rely on opinions posted on the Internet. However, unfamiliar places and rarely visited sightspots often do not accumulate sufficient number of valuable opinions. In this study, we propose a method that assigns most probable rating score to sightspots opinions discovered automatically by analyzing information collected from geotagged tweets. This is realized as an effort for complementing the lack of information for the least reviewed sightspots. Our method automatically assigns score to the extracted tweets regarding the target spot by learning contextual orientation of tweets with a state-of-the-art deep learning based approach. To develop the method, we firstly deployed three annotators who carefully annotated extracted location-clustered geotagged tweets for the training purposes. Next, by analyzing the semantic orientation of the tweets, we assign positive, neutral or negative sentiment to the tweets. Our method used a fine-tuned BERT model which analyses relative contextual orientation of tokens. The BERT model achieved an average F-score of 0.74, suggesting that, applying a state-of-the-art deep learning-based approach was useful in solving this task. The extracted tweets and generated score are then mapped on the designed Park Interface Review Site System as supplementary reviews for travelers seeking additional information about the related sightseeing spots.

Key Words *: Reviews, Score generation, Twitter, BERT*

1. INTRODUCTION

The Internet today, has vastly altered the data landscape, by accumulating large amounts of information. People, businesses, and devices have all become data factories that are pumping out large amounts of information to the Web each day Askitasklaust et al [1]. This huge amount of data shared on the Internet can be enhanced to foster various activities of a given specific area. A good example is tourism activities. Internet users can easily express their opinions about a product, service or a place they have recently visited using social media or review sites and reach millions of other potential visitors. In this way, people tend to transmit their daily events in the form of diaries and short sentences using online social services such as blogs, online posts, microblogs and other SNS. Among many Social Networking Services, the one that recently became popular for people to express their opinions, share their thoughts, and report real-time events, has been Twitter *1. With more than 336 million monthly active users and 500 million tweets sent daily*2, many companies and organizations became interested in utilizing data appearing on Twitter to study the opinions of people towards different products, services, facilities, and events taking place around the world.

Through Twitter, a great number of messages (known as tweets) are posted daily because of its simplicity. More-over, with GPS technology implemented in mobile phones and computers, sightseers often prefer to share their views and pictures regarding their tour experiences on Twitter. This type of information is valuable and important in fa-cilitating tourism activities of the specific area tagged with GPS information. Online opinions thus can have a great

*1 https://twitter.com/

*2 https://www.statista.com/topics/737/twitter impact on brand, product or place reputation. It is be-cause of this fact, some potential visitors make informed decisions based on online opinions. Primarily, there is a number of online review sites for tourism related activities. One example is TripAdvisor *3. But recently the problem of spam reviews has also been noticed on review sites and travel-related blogs*4. Fake reviews are also published on review sites to gain more customers, competition advantage or, on the other hand, decreases the popularity of business. This has become a serious problem due to spreading mis-information. Because of this problem, online review sites are slowly becoming less reliable[2].

Therefore, in this study, we propose a method that auto-matically adds opinion scores to the sightspots by learning to classify information about the target spot from extracted location-clustered geotagged tweets by using a deep learn-ing technique. Our goal is to supplement reviews to the least reviewed sightspots by extracting online tweets. Tweets are considered geotagged if they include the geolocation information and the name of the target spot. Tweets that include the name of the target spot are treated as a possible sightseers' opinions. The challenge in realizing this task, was to confirm the authenticity of the extracted tweets. To address this, we location-clustered the extracted tweets and later deployed three annotators who first annotated the tweets as either on-spot or not, and further calculated the kappa coefficient between the annotators as reported in our previous work [3]. Furthermore, annotators assign score to the tweets as either positive, neutral or negative. Finally, the assigned scores are mapped in the designed Park Inter-face Review Site System as opinions from the sightseers and current points of interest respectively. In this paper, we report on the performance of the constructed classifier and implementation of the proposed method.

In this study, we selected Serengeti and Ngorongoro Na-tional Parks in Tanzania as the target spots for our case study. The area covers more than twenty thousands square kilometers with many sighspots scattered in the area. Be-cause of its variety, some spots are unfamiliar to sightseers and therefore rarely visited, and therefore accumulate few reviews. Results published in this paper, represent an effort to complement reviews information for those areas. This study intends to support the local tourism sector in Tanza-nia specifically in the area of wildlife-based tourism as one of the promising and fastest-growing sector among others in Tanzania, with the selected target spot attracting the most sightseers [4] [5] [6].

The rest of this paper has been divided into five sec-tions as follows. Section 2. introduces related works and previous research. Section 3. describes the applied data. Experiment and analysis of the results are discussed in sec-tion 4.. In section 5. we present conclusions and future works.

2. RELATED WORK

Recently, various studies have been conducted on the analysis, extraction, provision and presentation of tourism-related information on the Web.

Oku et al [7] proposed a method of mapping geotagged tweets to tourists spots based on the substantial activity regions of spots.Their method learns from a one-class Sup-port Vector Machine (SVM) classifier which first extracts temporal and phrasal features of the pattern sentence and further map the tweets into respective regions. Location-based Social Networking Services such as Foursquare was useful in this study by providing geotagged message data. Shimada et al. [8] suggested a method to identify on-site likelihood adequateness of posted tweets through a two-stage method, rule-based filtering and a machine learn-ing technique. In their method, previous and next tweets were taken into consideration as a potential target defin-ing context information. The analysis of the experimental results shows the effectiveness of the applied combined techniques. Okamura et al [9] proposed an automatic score generation method to support local restaurants with small number of reviews by analyzing the reviews posted on the internet. Their method used a Convolution Neural Network to learn the correspondence between the characteristics of the evaluation information appearing in the review and the evaluation score. Also, Aramaki et al. [10] proposed a ma-chine learning-based approach to extract influenza tweets from Twitter. In their proposed method, tweets that men-tion actual influenza patients are extracted by SVM clas-sifier. Cheng et al. [11] proposed a method of predicting user's location by focusing on the content of the tweet.Their method relies on the content of the tweet, classification of words in tweets with a strong local geo-scope and a lattice-based neighborhood smoothing model. Miyabe et al. [12] proposed a method of extracting tweets that relate to the target spot by generating a classifier with an n-gram based feature.

While many of these studies, focus on extraction of infor-mation using either rule-based approaches or simple ma-chine learning classifiers (e.g., SVM), we focus on extrac-tion of online opinions and assigning scores by adopting a state-of-the-art neural network-based architecture (BERT).

3. APPLIED DATA

This study uses geotagged tweets collected from Twitter which is one of the most popular social network services and allows its users to share real-time opinions. Dataset used in this study contains 1,273 geotagged tweets and was collected by Silaa et al. [3]. The data was collected within a period of eight months, from June 2019 through February 2020. Tweets were collected by searching for the

*3 https://tripadvisor.com/

*4 https://www.theguardian.com/travel/2016/jul/08/trouble-with-tripadvisor-summer-food-drink-marina-oloughlin

Fig.1 An example of a geotagged tweet.

keywords “ngorongoro" and “serengeti" which could occur anywhere in the tweet that was finally included in the dataset. Tweets were collected through a streaming process using an accredited API*5 for this specific task from Twitter developers platform. Tweets are regarded explicitly geo-tagged if they include geocode information and the name of the target spot. To confirm authenticity of extracted tweets, geotagged tweets were clustered considering loca-tion using a K-means algorithm [13], and furthermore clas-sified by adopting a Support Vector Machine as a baseline approach and a fine-tuned pre-trained BERT neural lan-guage model.Three annotators were deployed to annotate location-clustered tweets with the kappa coefficient of 0.41 between annotators being recorded. Our classifier was then trained to predict weather a tweet was likely to be on-spot or not.

Table 1 Examples of annotated tweets.

Tweet	label	remarks
i’m at serengeti national park	on-spot	target spot
for booking safari serengeti, ngorongoro	not	advert

Table 2 A summary of Dataset used in this study.

Classification	on-spot	not
Tweets Counts	974	299

A set of 974 tweets were classified as on-spot while the remaining 299 tweets were classified as not on-spot and consisted mostly of advertisement tweets or tweets from different target spots. They were termed as irrelevant

*5 https://developer.twitter.com/en/apps/16398236

Table 3 Annotation summary - 3 scale range

Rating	3	2	1
Tweets Counts	664	276	34

Table 4 Annotation summary - 5 scale range

Rating	5	4	3	2	1
Tweets Counts	224	440	276	29	5

and therefore, this study uses the selected set of on-spot classified tweets (974 tweets).

a) Annotation

Annotation is a methodology of adding specific labels to a document. Manual text annotation is essential part of text mining. Automated text mining relies on manually annotated data by building their heuristic or statistical rules or neural networks on the basis of such annotated data [14]. In the annotation process, we define the text to annotate, set labels to put in tweets and set rules on how to deal with tweets that contain a certain degree of ambiguity. To accomplish this task, we asked three annotators who carefully annotated the tweets. We defined a rating scale for annotation with a different evaluation range, as follows.

Three scale interval

- 3 star - for a positive sentiment
- 2 star - for a neutral sentiment
- 1 star - for a negative

sentiment Five scale interval

- 5 star - for a very positive sentiment
- 4 star - for a positive sentiment
- 4 star - for a neutral sentiment
- 2 star - for a somewhat negative sentiment
- 1 star - for a harsh or negative sentiment

After annotation, we decided on the score information by taking the average score between three annotators for each specific tweet. Moreover, if a tweet received the same score from two different annotators, we used that annotated score.

4. EVALUATION EXPERIMENT

We fine-tuned the pre-trained BERT neural language model and used it for the classification task. BERT stands for Bidirectional Encoder Representation from Transform-ers. It is designed to pre-train deep bidirectional word representations from unlabeled text by jointly conditioning on both left and right context [15].

Annotated tweets were pre-processed by lowercasing, Non-Ascii letters, URLs and retweets were also removed.

Table 5 Example of 3-scale annotated tweets.

Tweet content	score
amazing elephant experience with #oliviatravel today in the #serengeti #grateful #elephants #wildlife at four season	3
ngorongoro crater at ngorongoro national park	2
worse places to get some writing done #amwritingscifi#tanzania #travel #writersofinstagram at kiota camp serenget	1

Table 6 Results summary.

Score range	Accuracy	F1
5-star score range	0.69	0.66
3-star score range	0.77	0.74

No lemmatization was perfomed. No stop-words were removed for fluency purposes. The original BERT-Base un-cased model has 12 transformer layers, 12 self-attention heads, 768-hidden size, and 110 million parameters. Specifically, an improved distilbert-base-uncased model*6 was used in this experiment. We deployed a ktrain library*7 which is a lightweight wrapper for tf.keras in Tensorflow 2 - a framework for neural net-works. Finally, we trained our model in consecutive 2 epochs.

a) Discussion

As observed in Table 6 the prediction performance was better for the 3-star range interval compared to 5-star score range. A 3-score range setup outperform a five-score range scale with an F-score of 0.74.This happened most probably due to data imbalance between both scenarios, 5-star and 3-star, and the smaller number of classes results in more samples per each class and eventually allows for better generalization of data. There was classifier misjudgement also between annotated score and predicted score as shown with tweets example in table 7. Fig 2 also shows our model evaluated a negative sentiment tweet (tweet number 16) as positive sentiment. One way to improve our model performance is to remove tweets with high degree of ambiguities in training set.This will be our consideration in our next experiment. The results demonstrated that our proposed method, although not ideal, can be used for score generation.

Therefore, a 3-star score range was adopted for application in the designed system expressing weather a tweet contains a positive, neutral, or negative review of the sightspot.

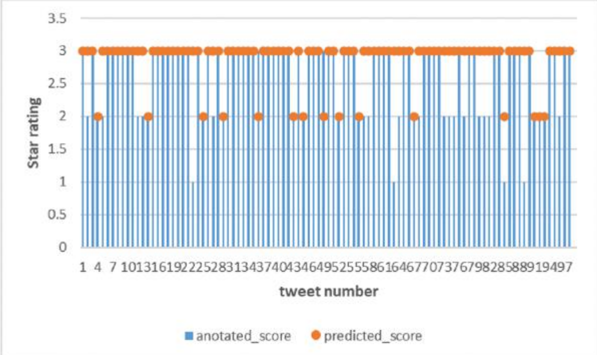

Fig.2BERT score prediction results for a 3-class range

b) Park Review Site Navigation System

The results of this study, namely, the extracted tweets with their generated opinion scores, were therefore mapped in the designed system as supplementary reviews.

The designed system*8 has a database which holds more than 200 sightspots available in the target spots and displays internet opinions with rating score of various sightspots registered in the database as per user selection. In addition, the system is using Google Maps API (as observed from the interface page) in figure 3, with a function to display routes from point to point in the target spot. Also, user can search for sighspots, read extracted reviews from tweets and use it for navigation.

5. CONCLUSIONS

In this paper, we focused on assigning opinion scores to the extracted tweets by using a BERT neural language model-based classifier which learns semantic orientation of the geotagged tweets. Discovered tweets are further mapped in the designed review system as supplement reviews from the internet as an effort to provide opinions for sighspots having insufficient reviews. BERT achieved a 0.74 F-score measure for a three score range scale suggest-ing it can be used for further implementations.

BERT model trained specifically on Twitter was recently released which has an advantage of better coverage of the

*6 https://huggingface.co/distilbert-base-uncased

*7 https://github.com/amaiya/ktrain/

*8 http://kcloss.cs.kitami-it.ac.jp/kcloss/WebAppDctoss2/

Table 7 Examples of misjudged tweets

Tweet content	anotation	prediction
we spent our final safari day at ngorongoro crater it was surprisingly cold but we had a rare chance of	1-star	3-star
there is no #wifi on a #safari but youll find a better connection #tanzania #ngorongoro at ngorongoro	1-star	3-star

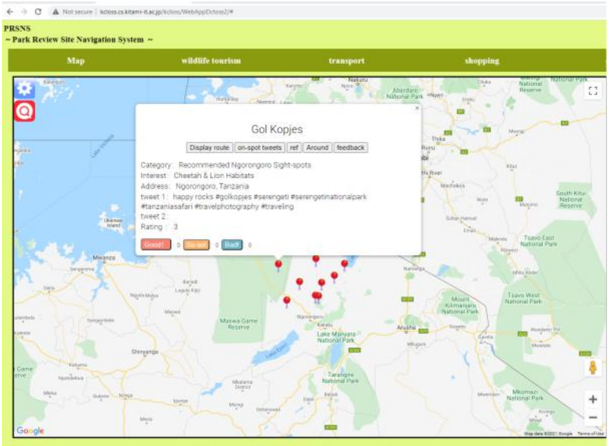

Fig.3 System interface.

vocabulary used on the SNS. Therefore in the future, we plan to improve the the prediction results by adopting it. In additional, we plan to apply different pre-processing tech-niques to the dataset, example; lematization, part of speech tagging, stop words and further analyze the performance of our model.Moreover,In this study a Distilled BERT model was used, but different pre-trained BERT models exist. It will be interesting to use them and further compare the ef-ficacy.

REFERENCES

1) Nikolaos AskitasKlaus and F. Zimmermann. The in-ternet as a data source for advancement in social sci-ences. International Journal of Manpower on "Using internet activity data to analyze human resources is-sues". Discussion paper series, 2015.
2) Reza Farahbakhsh Noel Crespi Saeedreza Shehnepoor, Mostafa Salehi. Netspam: A network-based spam de-tection framework for reviews in online social media. *IEEE Transactions on Information Forensics and Se-curity*, Vol. 12, No. 7, 2017.
3) Michal Ptasyznski Victor Silaa, Fumito Masui. The-sis: Development of a method of automatic rating and classification of local sightspots by learning from geo-tagged tweets. 2020.
4) NBS. 2016 tanzania in figures. In *The United Republic of Tanzania*, 2017.
5) M. Makanjio Okello and S. Yerian. Tourists satis-faction in relation to attractions and implications for conservation in the protected area in the northern cir-cuit, tanzania. In *Journal of sustainable Tourism Vol-ume 17, 2009 - Issue 5,Pages 605-625*. Informa UK Limited, 2009.
6) SLS Maganga RB Modest FS Salehe SB Mariki, SN Hassan. Wildlife-based domestic tourism in tan-zania: Experience from northern tourist's circuit. In *Ethiopian Journal of Environmental Studies and Man-agement Vol. 4 No.4 2011*. AFRICAN JOURNALS ONLINE (AJOL), 2011.
7) Fumio Hattori Kenta Oku. Mapping geotagged tweets to tourist's spots considering activity region of spot. In *Intelligent Systems Reference Library 90, Tourism Informatics Towards Novel Knowledge Based Ap-proaches*. Springer, 2015.
8) Shunsuke Inoue Tsutomu Endo . Kazutaka Shimada, Yurie Onitsuka. On-site likelihood identification of tweets using a two-stage method. In *Intelligent Systems Reference Library 90, Tourism Informatics Towards Novel Knowledge Based Approaches*. Springer, 2015.
9) Michal Ptaszynski Hiroshi Masui Kamemaru Toshikazu Yasunari maeda Eigi Kuroda Shin Oka-mura, Fumito Masui. Proposal of automatic evalu-ation score generation method by restaurant reviews. In *The international Workshop on modern science and Technology*, 2018.
10) Morita.M Aramaki.E, Maskawa.S. Twitter catches the flu:detecting influenza epidemics using twitter. In *In proceedings of Conference on Empirical Methods in Natural Language Processing(EMNLP 2011)*. Asso-ciation for Computational Linguistics, 2011.
11) Lee.K Cheng.Z, Caverlee.J. You are where you tweet:a content-based approach to geo locating twitter users. In *In proceedings of the 19th ACM International con-ference on Information and Knowledge Management*. ACM Digital Library, 2010.
12) Kubo.K Aramaki.E Miyabe.M, Kita.Y. Extracting as-pect record related to a location from microblog(in japanese). In *In proceeding of the 20th Annual meeting of the Association for Natural language processing,pp.420-423*, 2014.
13) M. A. Wong J.A. Hartigan. Algorithm as 136: A k-means clustering algorithm. In *Journal of the Royal statistical Society, Series C (Applied statistics*. John Wiley Sons, Inc., 1979.

14) Marina Sokolova Victoria Bobicev. Inter-annotator agreement in sentiment analysis:machine learning per-spective. In *In Proceedings of Recent Advances in Nat-ural Language Processing, pages 97–102*. INCOMA Ltd., 2017.
15) Kenton Lee Kristina Toutanova Jacob Devli, Ming-Wei Chan. Bert: Pre-training of deep bidirectional transformers for language understanding. In *In Proceedings of NAACL-HLT 2019, pages 4171–4186*. As-sociation for Computational Linguistics, 2019.