شبكة المعلومات الجامعية

التوثيق الإلكتروني والميكروفيلم

بسم الله الرحمن الرحيم

HANAA ALY

شبكة المعلومات الجامعية

التوثيق الالكتروني والميكروفيلم

HANAA ALY

شبكة المعلومات الجامعية

التوثيق الإلكتروني والميكروفيلم

جامعة عين شمس

التوثيق الإلكتروني والميكروفيلم

قسم

نقسم بالله العظيم أن المادة التي تم توثيقها وتسجيلها
علي هذه الأقراص المدمجة قد أعدت دون أية تغيرات

يجب أن

تحفظ هذه الأقراص المدمجة بعيدا عن الغبار

HANAA ALY

Computer Science Department
Faculty of Computer and Information Sciences
Ain Shams University

Sentiment Analysis in Tourism

Thesis submitted as a partial fulfilment of the requirements for the degree of Master of Computer Science

By

Sarah Osama Anis
Demonstrator at Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University.

Under the Supervision of

Prof. Dr. Mostafa Aref
Professor of Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University.

Dr. Sally Saad Ismail
Lecturer of Computer Science Department, Faculty of Computer and Information Sciences, Ain Shams University

2021

Acknowledgments

First and foremost, praises and thanks to the God, the Almighty, for his showers of blessings throughout my journey to complete the research successfully.

I would like to express my deep and sincere gratitude to my research supervisors, Prof. Mostafa Aref and Dr Sally Saad for giving me the opportunity to do this research and providing me with invaluable guidance throughout this research. Their support, vision and motivation have deeply inspired me. It was a great privilege and honor to work and study under their guidance.

I am extremely grateful to my family for their constant love, support, prayers and sacrifices for educating and preparing me for my future. I am also so thankful to my beloved husband for his love, understanding, prayers and continuous support to complete this research work. And my dear son, who served as an inspiration to pursue my dreams.

Many thanks and appreciations also go to my committee members and everyone who played a role in my academic accomplishments. Thank you all for your unwavering support. Without you, I could never have reached this current level of success.

Abstract

Sentiment Analysis is an automated process of analysing people's opinions and feelings using Natural Language processing tools. As everything is shifting online, the demand for sentiment analysis has increased tremendously. In tourism, Sentiment analysis can help to comprehend tourists concerns and complaints which will benefit organizations in this field with accurate sentiment tracking of their customers, enabling them to improve customer experience. Tourism-related websites have turned into an incredible data source that impacts the tourism industry from many points of view. Tourists express their opinions regarding products and services online daily. The interest in understanding and analysing customer opinions has increased significantly over the past few years as it is vital for the decision making of both customers and companies. Sentiment analysis is the practice of applying natural language processing, statistics and machine learning methods to extract and identify the common opinion behind the text in a review, blog discussion, news, comments or any other document. Sentiment analysis has great potential to directly understand tourists' opinions.

This thesis tackles a comprehensive overview of the latest update in this field giving a nearly full image of sentiment analysis approaches, techniques, and challenges in analysing the correct meaning of sentiments and detecting the suitable sentiment polarity in the field of tourism. It discusses the general process of sentiment analysis with its stages along with recent studies reviewed for each stage. It gives a detailed description of the main sentiment classification approaches which are machine learning approach, Lexicon-based approach and Hybrid approach. A comparative analysis between the sentiment classification approaches, methods and techniques is also presented to highlight the differences between approaches and the advantages and disadvantages of each approach and technique. There are several challenges in sentiment analysis that are highlighted in this thesis that help shed light on areas that are less investigated in this field. Recommendations to solve these challenges are also presented.

In this thesis, an approach is introduced that automatically perform sentiment analysis for hotel reviews provided by customers from one of the leading travel sites. Different techniques were investigated, Fuzzy C-means clustering algorithm was used for sentiment detection to extract subjective sentences from objective ones. Sentiment detection could be viewed as a prior step to increase the accuracy of sentiment classification. Sentiment detection is an important sub-task of sentiment analysis that can prevent a sentiment classifier from considering the deceptive or misleading text in online reviews. Sentiment classification determines the overall polarity of opinion whether it's positive or negative. For sentiment classification, hotel reviews have been analysed using various techniques like Naïve

Bayes, K-Nearest Neighbour, Support Vector Machine, Logistic Regression, and Random Forest. An ensemble learning model was also proposed that combines the five classifiers. Ensemble learning was used in order to achieve better results, as it is commonly known to outperform the performance of single classifiers. We have also investigated the importance of deep learning in sentiment analysis and its ability to improve the accuracy of the sentiment prediction. We have proposed a deep learning approach based on word embedding and gated recurrent unit to solve the sentiment classification problem. Finally results of each classifier were compared. Ensemble classifier achieved 86.2% accuracy and best results of the five classifiers were obtained by the Support Vector Machine with 86.3% accuracy. Our deep learning approach outperformed the performance of other methods with accuracy 89% and 92% F-score.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ABSA	Aspect-Based Sentiment Analysis
ANFIS	Adaptive Neuro-Fuzzy Inference System
CNN	Convolutional Neural Network
DT	Decision Tree
ELM	Extreme Learning Machine
FCS	Fuzzy Control System
FP	False Positive
FN	False Negative
GRU	Gated Recurrent Unit
HS-SVM	Hybrid Scoring Support Vector Machine
HTML	Hypertext Mark-up Link
KNN	K-Nearest Neighbour
LMSOR	Language Model based Supervised Opinion Retrieval
LMSSOR	Language Model based Semi-Supervised Opinion Retrieval
LSA	Latent Semantic Analysis
LSTM	Long Short-Term Memory
ML	Machine Learning
MLP	Multi-Layer Perceptron
NB	Naïve Bayes
PCA	Principal Component Analysis
POS	Part Of Speech
RP	Random Projection
SA	Sentiment Analysis
SO	Sentiment Orientation
SVM	Support Vector Machine
UGC	User Generated Content

List of Publications

- Sarah Anis, Sally Saad, Mostafa Aref,"Sentiment Analysis of Hotel Reviews Using Machine Learning Techniques". In Proceedings of the International Conference on Advanced Intelligent Systems and Informatics. AISI 2020. Advances in Intelligent Systems and Computing, vol 1261, pp 227-234. Springer, Cham, 2020.

- Sarah Anis, Sally Saad, Mostafa Aref, "A survey on sentiment analysis in tourism", International Journal of Intelligent Computing And Information Sciences (IJICIS) , Cairo, Egypt, 2020.

- Sarah Anis, Sally Saad, Mostafa Aref, "Deep Learning-Based Approach for Sentiment classification of Hotel Reviews", the 3rd International Conference on Communication and Computational Technologies (ICCCT 2021), Springer, Souvenir, India, Feb 2021.

Chapter 1

Introduction

Chapter 1 : Introduction

Nowadays, people generally prefer to communicate and socialize on the web. With the widespread usage of social media in our daily lives, social media websites became a vital and major source of data about user reviews in various fields. Sentiment classification is the task of classifying the data into some categories mainly to positive or negative opinions that have been expressed on a certain product, organization, or event [1]. Opinions of users in social media are generally expressed by using informal, slang, and non-standard words, which increase the complexity of the sentiment classification process [2].

Customer reviews on social media often reflect joy, dissatisfaction, frustration, happiness, and different sentiments. Taking advantage of these huge volumes of subjective information is of great value to tourism associations and organizations which aim to increase profitability and enhance or maintain customer satisfaction. Sentiment polarity classification can be binary, ternary or ordinal classification [3]. In binary classification, the polarity of a given review is classified as positive or negative assuming that the text is subjective in the first place. Sentences with subjective expressions includes opinions, beliefs, personal feelings and views while objective expressions include facts, evidences and measurable observations [4]. The assumption made that reviews are subjective is not necessarily true; customer reviews provided through the text is considered either subjective or objective which means a ternary classification that requires the third category is needed. Recently, most of the sentiment analysis approaches apply sentiment detection first which differentiate between objective and subjective reviews using a binary classifier then determines the sentiment polarity of subjective reviews using a binary polarity classifier. There are many challenges that face the sentiment analysis process In addition to the binary and ternary classification, ordinal classification can be employed to automatically rate reviews based on an ordinal regression model by the means of a numbered ranking scale [5, 6]. Sentiment analysis can be performed at word, sentence, paragraph or document levels. In document-level, the whole document is assigned a single sentiment while in paragraph-level sentiment analysis, the sentiment of each paragraph is determined separately. Other approaches consider applying sentiment analysis on each sentence or even each word individually. It is more challenging to accurately extract polarity in sentence-level since sentences contain a small number of words compared with paragraphs and documents.

1.1 Motivation

The domain of tourism extended activity online in the most recent decade. There are many individuals that book accommodation online everyday as it is less tedious, less expensive and they can get a point by point data about facilities and location of hotels. The benefit of having fast access to information and feedback, make users lean towards internet booking. Studies about consumers' online behaviour