A Mini Project Report On

Sentiment Analysis of Twitter data Based on Ordinal Regression

A thesis

submitted in partial fulfillment of the requirement for

the award of the degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

by

P.LAKSHMI SATYA

19021D0520

Under the Supervision of

Smt. S.S.S.N USHA DEVI N

Assistant Professor

CSE,UCEK

JNTUK, Kakinada.

Submitted to

UNIVERSITY COLLEGE OF ENGINEERING KAKINADA

JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY KAKINADA

KAKINADA - 533003

(2019-2021)

UNIVERSITY COLLEGE OF ENGINEERING KAKINADA
JAWAHARLAL NEHRU TECHNOLOGICAL UNIVERSITY KAKINADA

C E R T I F I C A T E

This is to certify that the thesis entitled **"Sentiment Analysis of Twitter data Based on Ordinal Regression"** being submitted by **PAMULAPATI LAKSHMI SATYA** bearing the Roll number **19021D0520** in the partial fulfilment for the award of degree of **MASTER OF TECHNOLOGY** in **COMPUTER SCIENCE AND ENGINEERING** to UCEK, JNTUK, Kakinada, Andhra Pradesh, India, is a record of bona fide work carried out by her under the guidance and supervision of the Department.

Internal Guide

Smt. S.S.S.N USHA DEVI N

Assistant Professor

CSE,UCEK

JNTUK, Kakinada.

Head of the Department

Dr. D. HARITHA

Professor & HOD

CSE,UCEK,

JNTUK, Kakinada

External Examiner

Declaration

The MINI PROJECT work entitled **"SENTIMENT ANALYSIS OF TWITTER DATA BASED ON ORDINAL REGRESSION"** has been carried out by me at **Jawaharlal Nehru Technological University, Kakinada** under the Guidance and Supervision of **Smt. S.S.S.N USHA DEVI N, UCEK**, JNTUK, Kakinada.

This work is original and has not been submitted to any other University / Institute for the award of any degree or diploma.

Pamulapati Lakshmi Satya

(19021D0520)

Acknowledgement

I wish to express my sincere gratitude to my project guide **Smt. S.S.S.N USHA DEVI N, UCEK** for providing me with an opportunity to do my project on "**SENTIMENT ANALYSIS OF TWITTER DATA BASED ON ORDINAL REGRESSION"**.

I am grateful for giving me valuable advice and constant source of encouragement throughout the project.

Without her valuable inputs and continuous suggestions, this project would not have been possible.

Finally I thank my family and friends for their support.

Pamulapati Lakshmi Satya

ABSTRACT

Sentiment analysis (also known as opinion mining or emotion AI) refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information.

On internet, Opinion (sentiments on topic) mining is helping users in knowing the quality of any organization or products, if any user has good experience on any product or company then he will express good reviews/opinion and by seeing this opinion others users can know the quality of the product.

In today's online social networks like twitter all peoples are expressing their opinions and social networking sites are developing new techniques to detect sentiments from this opinions.

Ordinal regression is a statistical technique that is used to predict behaviour of ordinal level dependent variables with a set of independent variables. The dependent variable is the order response category variable and the independent variable may be categorical or continuous.

We propose 5 levels of sentiments detection such as High Positive, Moderate Positive, Neutral, High Negative and Moderate Negative. To detect sentiments we are using two Ordinal Regression Machine Learning algorithms such as Decision Tree and Random Forest.

In this project, we give training set tweets as input and classifier predicts sentiments of test set by using all independent words from this tweets.

The project consists of first pre-processing tweets and using a feature extraction method that creates an efficient feature. Then, under several classes, these features scoring and balancing. Decision Trees (DTs) and Random Forest (RF) algorithms are used for sentiment analysis classification. For the actual implementation of this system, a twitter dataset publicly made available by the NLTK corpora resources can be used.

TABLE OF CONTENTS

Chapter 1. INTRODUCTION

1.1 General preview

With the rapid development of social networks and micro blogging websites, Micro blogging websites have become one of the largest web destinations for people to express their thoughts, opinions, and attitudes about different topics [. Twitter is a widely used micro blogging platform and social networking service that generates a vast amount of information.

In recent years, researchers preferably made the use of social data for the sentiment analysis of people's opinions on a product, topic, or event. Sentiment analysis, also known as opinion mining, is an important natural language processing task. This process determines the sentiment orientation of a text as positive, negative, or neutral .

Twitter sentiment analysis is currently a popular topic for research. Such analysis is useful because it gathers and classifies public opinion by analysing big social data. However, Twitter data have certain characteristics that cause difficulty in conducting sentiment analysis in contrast to analysing other types of data.

Tweets are restricted to 140 characters, written in informal English, contain irregular expressions, and contain several abbreviations and slang words. To address these problems, researchers have conducted studies focusing on sentiment analysis of tweets .

Twitter sentiment analysis approaches can be generally categorized into two main approaches, the machine learning approach, and a lexicon-based approach. In this project, we use machine learning techniques to tackle twitter sentiment analysis.

Most classification algorithms are focused on predicting nominal class data labels. However, a rule for predicting categories or labels on an ordinal scale involves many pattern recognition issues. This type of problem, known as ordinal classification or ordinal regression. Recently, ordinal regression has received considerable attention.

Ordinal regression issues in many fields of research are very common and have often been regarded as standard nominal problems that can lead to non-optimal solutions.

In fact, Ordinal regression problems with some similarities and differences can be said to be between classification and regression. Medical research, age estimation, brain-computer interface, face recognition, facial beauty evaluation, image classification, social sciences, text classification, and more are some of the fields where ordinal regression is found.

Some studies suggest using machine learning techniques to solve regression problems to improve the sentiment analysis classification of Twitter data performance and predict new results. The main advantage of this method is the achievement of improved results.

1.2 Scope

The current study mainly focuses on the sentiment analysis of Twitter data (tweets) using different machine learning algorithms to deal with ordinal regression problems. In this project, we propose an approach including pre-processing tweets, feature extraction methods, and constructing a scoring and balancing system, then using different techniques of machine learning to classify tweets under several classes.

1.3 Objectives

- To classify the training set tweets into five classes such as High Positive, Moderate Positive, Neutral, High Negative and Moderate Negative.
- To predict the Sentiments of test set tweets using Supervised Machine Learning algorithms such as Decision tree and Random Forest.

Chapter 2. REVIEW OF LITERATURE

B. O'Connor, R. Balasubramanyan, B. R. Routledge and N. A. Smith in the paper "**From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series" [1]** discussed that the extremely simple text analysis techniques are demonstrated to correlate with polling data on consumer confidence and political opinion, and can also predict future movements in the polls . We analyze several surveys on consumer confidence and political opinion over the 2008 to 2009 period, and find their correlation to sentiment word frequencies in contemporaneous Twitter messages. While our results vary across datasets, in several cases the correlations are as high as 80%, and capture important large-scale trends. The results highlight the potential of text streams as a substitute and supplement for traditional polling.

M. A. Cabanlit and K. J. Espinosa in the paper "**Optimizing N-gram based text feature selection in sentiment analysis for commercial products in Twitter through polarity lexicons" [2]** aims to optimize N-gram based text feature selection in sentiment analysis for commercial products in twitter through polarity lexicons. This can be done by merging dictionary-based weighing with naïve-Bayes classification of sentiments. The study is still ongoing but partial results show potential.

S.-M. Kim and E. Hovy in the paper titled **"Determining the sentiment of opinions" [3]** said that identifying sentiments (the affective parts of opinions) is a challenging problem. They present a system that, given a topic, automatically finds the people who hold opinions about that topic and the sentiment of each opinion. The system contains a module for determining word sentiment and another for combining sentiments within a sentence. The basic approach is to

assemble a small amount of seed words by hand, sorted by polarity into two lists—positive and negative—and then to grow this by adding words obtained from WordNet. We experiment with various models of classifying and combining sentiment at word and sentence levels.

P. Jain and P. Dandannavar, in the paper **"Application of machine learning techniques to sentiment analysis"[4**] discussed in detail various steps for performing sentiment analysis on twitter data using machine learning algorithms. Once an appropriate dataset is collected, the next step is to perform preprocessing on data (tweets) by using NLP based techniques, followed by feature extraction method in order to extract sentiment relevant features. Finally, the model is trained using machine learning classifiers like Naïve Bayes, Support Vector Machines or Decision trees and is tested on test data. The performance of the model is measured in terms of accuracy, precision, recall and F-score. This work performed analysis on datasets of different sizes and domains to demonstrate that the proposed framework works on data of all sizes and domains. This work uses Apache Spark to obtain the accurate results fast.

A. K. Jain, R. P. W. Duin and J. C. Mao in the paper **"Statistical pattern recognition: A review"[5]** summarizes and compare some of the well-known methods used in various stages of a pattern recognition system for new and emerging applications, such as data mining, web searching, retrieval of multimedia data, face recognition, and cursive handwriting recognition. The four best known approaches for pattern recognition which include template matching, statistical classification, syntactic or structural matching, and neural networks are discussed.

P. A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernández-Navarro and C. Hervás-Martínez in the paper **"Ordinal regression methods: Survey and experimental study"[6]** discussed about difference between ordinal regression problem and nominal classification problem and the taxonomy of ordinal regression methods is proposed, dividing them into three main groups: naive approaches, binary decompositions and threshold models. Furthermore, the most important methods of Naive approach family which includes Regression. Nominal classification, cost sensitive classification and different Threshold models which include Support Vector model, Perceptron learning, Discriminant Learning are discussed.

Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua in the paper **''Ordinal regression with multiple output CNN for age estimation" [7]** discussed about the End-to-End learning approach to address ordinal regression problems using deep Convolutional Neural Network, which could simultaneously conduct feature learning and regression modeling. In particular, an ordinal regression problem is transformed into a series of binary classification sub-problems. And a multiple output CNN learning algorithm was proposed to collectively solve these classification sub-problems, so that the correlation between these tasks could be explored. A new age dataset is released to the community for age estimation, which is the largest public dataset to date.

X. Chen, M. Vorvoreanu and K. Madhavan in the paper **"Mining social media data for understanding students learning experiences" [8]** developed a workflow to integrate both qualitative analysis and large-scale data mining techniques. Focus is on engineering students Twitter posts to

understand issues and problems in their educational experiences. A qualitative analysis is first conducted on samples taken from about 25,000 tweets related to engineering students college life. Engineering students encounter problems such as heavy study load, lack of social engagement, and sleep deprivation. Based on these results, a multi-label classification algorithm was implemented to classify tweets reflecting students problems. Then the algorithm is used to train a detector of student problems from about 35,000 tweets streamed at the geo-location of Purdue University. This paper, for the first time, presents a methodology and results that show how informal social media data can provide insights into students' experiences.

V. Singh and S. K. Dubey, **"Opinion mining and analysis: A literature review"[9]** says that the sentiment analysis is done on the social issues and events taking data from the social sites. Sentiment analysis has a wide variety of application in summarizing reviews, classifying reviews, information system, market analysis and decision making. Sentiment analysis is a broad range of fields of natural language processing and text mining. It is found that different types of features and classification techniques are combined in an efficient way to enhance the sentiment classification. The data is collected from the social sites, blogs and micro blogging sites, and effectively perform sentiment analysis on the data using various machine learning techniques.

N. R. Kasture and P. B. Bhilare in the paper **"An approach for sentiment analysis on social networking sites"[10]** discussed an approach for extraction of the sentiment on widely used social networking sites. The approach used is the logical approach which mainly focuses on combinatory categorical

grammar, lexicon acquisition and annotation and semantic networks for analyzing the sentiments of the text. Focus on the formal logical approach is motivated by some issues with reference to machine learning. Compared to machine learning approach, formal logical systems are extremely precise in results. Formal logical approach to a certain extent solves the assignment of polarity values on lexical level quite well and another advantage is that it is loosely coupled to domain. A logical method also suffers from some issues like robustness, e.g. if there are missing, or incorrect axioms, the system will not work as desired. There is a scope of improvement in the suggested approach to resolve such issue.

L. Breiman in the document ‘**‘Random forests” [11]** said that Random forests are an effective tool in prediction. Because of the Law of Large Numbers they do not overfit. Injecting the right kind of randomness makes them accurate classifiers and regressors. Furthermore, the framework in terms of strength of the individual predictors and their correlations gives insight into the ability of the random forest to predict. Using out-of-bag estimation makes concrete the otherwise theoretical values of strength and correlation. Forests give results competitive with boosting and adaptive bagging, yet do not progressively change the training set. Their accuracy indicates that they act to reduce bias.

A. Celikyilmaz, D. Hakkani-Tür, and J. Feng in the paper **‘‘Probabilistic model-based sentiment analysis of Twitter messages” [12]** presented a machine learning approach to sentiment classification on twitter messages (tweets). In this work each tweet is classified into two categories: polar and non-polar. Tweets with positive or negative sentiment are considered polar. They are considered non-polar otherwise. Data is collected from Twitter API and normalization is performed which include Text processing and Pronunciation

based clustering. Then LDA model is used to find grouping of polar words for identifying sentiment.

V. Cherkassky and F. M. Mulier in the book **"Learning From Data: Concepts Theory and Methods" [13]** discussed about Formulation of Learning Problem, Classification, Regression, Curse and Complexity of Dimensionality, Bias Variance trade-off, Linear Regression, Non-linear optimization methods, Methods for Data reduction and Dimensionality Reduction, Methods for Regression, Support Vector Machines. Noninductive inference and Alternative Learning formulations.

L. Li and H. Lin in the paper **"Ordinal regression by extended binary classification" [14]** presented a reduction framework from ordinal regression to binary classification based on extended examples. The framework consists of three steps: extracting extended examples from the original examples, learning a binary classifier on the extended examples with any binary classification algorithm, and constructing a ranking rule from the binary classifier. The framework is developed using perceptron kernel and it is tested with 3 binary classification algorithms which include Quinlan's C4.5, Ada Boost-stump and SVM with the perceptron kernel.

Within the three SVM-based approaches, the two with the perceptron kernel are better than SVOR-IMC with the Gaussian kernel in test performance. Direct reduction to the standard SVM performs similarly to SVOR-IMC with the same perceptron kernel, but is much easier to implement. In addition, direct reduction is significantly faster than SVOR-IMC in training

A. Go, R. Bhayani and L. Huang in the paper **"Twitter sentiment classification using distant supervision" [15]** introduced a novel approach for

automatically classifying the sentiment of Twitter messages. These messages are classified as either positive or negative with respect to a query term. This is useful for consumers who want to research the sentiment of products before purchase, or companies that want to monitor the public sentiment of their brands. The results of machine learning algorithms for classifying the sentiment of Twitter messages using distant supervision are presented. Training data consists of Twitter messages with emoticons, which are used as noisy labels. It is shown that machine learning algorithms (Naive Bayes, Maximum Entropy, and SVM) have accuracy above 80% when trained with emoticon data. This paper also describes the preprocessing steps needed in order to achieve high accuracy. The main contribution of this paper is the idea of using tweets with emoticons for distant supervised learning. The usage of unigrams, bigrams, unigrams and bigrams, and parts of speech are used as features.

Chapter 3. CONCLUSION

- A Comparative Study is made by analyzing the works done by the research community in the field of Sentiment Analysis.
- By analyzing their works, a methodology is proposed .In this project, Data is collected and Preprocessing is done, then features are extracted and then Sentiments of Twitter data are determined by building a balancing and scoring model, afterwards, tweets are classified into several ordinal classes using machine learning classifiers.
- Experimental results will be performed using Classifiers, such as Decision Trees and Random Forest. This approach can be optimized using Twitter data set that is publicly available in the NLTK corpora resources.

REFERENCES

1. B. O'Connor, R. Balasubramanyan, B. R. Routledge and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series", Proc. ICWSM, vol. 11, no. 129, pp. 1-2, 2010.
2. M. A. Cabanlit and K. J. Espinosa, "Optimizing N-gram based text feature selection in sentiment analysis for commercial products in Twitter through polarity lexicons", *Proc. 5th Int. Conf. Inf.* Intell. Syst. Appl. (IISA), pp. 94-97, Jul. 2014.
3. S.-M. Kim and E. Hovy, ''Determining the sentiment of opinions,'' in Proc. 20th Int. Conf. Comput. Linguistics, Aug. 2004, p. 1367.
4. P. Jain and P. Dandannavar, ''Application of machine learning techniques to sentiment analysis,'' in Proc. 2nd Int. Conf. Appl. Theor. Comput. Commun. Technol. (iCATccT), Jul. 2016, pp. 628–632.
5. A. K. Jain, R. P. W. Duin, and J. C. Mao, ''Statistical pattern recognition: A review,'' IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 1, pp. 4–37, Jan. 2000.
6. P. A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernández-Navarro, and C. Hervás-Martínez, ''Ordinal regression methods: Survey and experimental study,'' IEEE Trans. Knowl. Data Eng., vol. 28, no. 1, pp. 127–146, Jan. 2016.
7. Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, ''Ordinal regression with multiple output CNN for age estimation,'' in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 4920–4928.

8. X. Chen, M. Vorvoreanu, and K. Madhavan, ‘‘Mining social media data for understanding students’ learning experiences,’’ IEEE Trans. Learn. Technol., vol. 7, no. 3, pp. 246–259, Jul./Sep. 2014.
9. V. Singh and S. K. Dubey, ‘‘Opinion mining and analysis: A literature review,’’ in Proc. 5th Int. Conf.-Confluence Next Gener. Inf. Technol. Summit (Confluence), Sep. 2014, pp. 232–239.
10. N. R. Kasture and P. B. Bhilare, ‘‘An approach for sentiment analysis on social networking sites,’’ in Proc. Int. Conf. Comput. Commun. Control Autom., Feb. 2015, pp. 390–395..
11. L. Breiman, ‘‘Random forests,’’ Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.
12. A. Celikyilmaz, D. Hakkani-Tür, and J. Feng, ‘‘Probabilistic model-based sentiment analysis of Twitter messages,’’ in Proc. IEEE Spoken Lang. Technol. Workshop, Dec. 2010, pp. 79–84.
13. V. Cherkassky and F. M. Mulier, Learning From Data: Concepts, Theory, and Methods. Hoboken, NJ, USA: Wiley, 2007.
14. L. Li and H. Lin, ‘‘Ordinal regression by extended binary classification,’’ in Proc. Adv. Neural Inf. Process. Syst., 2007, pp. 865–872.
15. A. Go, R. Bhayani, and L. Huang, ‘‘Twitter sentiment classification using distant supervision,’’ Processing, vol. 150, no. 12, pp. 1–6, 2009.