RESEARCH ARTICLE OPEN ACCESS

Text and Emoji Classification - The Literature Survey

Vishwajit Jamdade*, Hrutika Salunkhe**, Vaibhav Kale**

*(Assistant Professor, Computer Engineering, SVPM's College of Engineering Malegaon(bk), Baramati
Email: rutujataware3@gmail.com)

**(UG Students, Computer Engineering, SVPM's College of Engineering Malegaon(bk), Baramati)
Email: jamdadevishwajit11@gmail.com,hrutikasalunkhe763@gmail.com, kale38182@gmail.com)

--************************----------------------------------

Abstract:

Elaboration of technologies and acceptance of social media tools and operations, new doors of occasion have been opened for using data logical in gaining meaningful perceptivity from unshapedinformation.In this design we are working on opinion mining that's classifying reviews according to the sentiment expressed in them positive, negative or neutral. There are number of online micro-blogging and social-networking platform which allows druggies to write short status updates. It's a fleetly expanding service with over 200 million registered druggies out of which 100 million are active druggies and half of them log on social media on a diurnal base- generating nearly 250 million post per day. Due to this large quantum of operation we hope to achieve a reflection of public opinion by assaying the Text and Emoji expressed in the post/ reviews. assaying the public reviews is important for numerous operations similar as enterprises trying to find out the response of their products in the request, prognosticating political choices and prognosticating socioeconomic marvels like stock exchange. The end of this design is to develop a functional classifier using machine learning for accurate and automatic Emoji bracket of reviews post.

Keywords **—Machine Learning, classification, Social Networking,Functional Classifier.**

--************************----------------------------------

I. INTRODUCTION

Text and emoji classification also known as Opinion Mining refers tothe use of natural language processing, text analysis to systematically identify, extract, quantify, and study affecting states and subjective information.Text and Emoji Classification is widely applied to reviews and survey responses, onlineand social media, and health-care materials for applications that range frommarketing to customer service to clinical medicine. In this project, we aim toperform Text and emoji classification of product based reviews. Data used inthis project are online product reviews data-set downloaded from Internet. Weexpect to do review-level categorization of review data with promising outcomes.

II. LITERATURE SURVEY

The paper "Sentiment Analysis of Tweets Including Emoji Data" is a study that explores the use of emoji in sentiment analysis of tweets. The study was conducted by researchers at the University of Warwick, UK and was published in the Journal of Web Science in 2017.

The goal of the study was to investigate the impact of Emoji on sentiment analysis and to determine whether incorporating emoji data into sentiment analysis models could improve their accuracy[1]. To do this, the researchers analyzed a data-set of tweets that included both textual content and emoji data.

The study found that incorporating emoji data into sentiment analysis models can improve their accuracy, particularly in cases where the textual content alone is ambiguous or difficult to interpret. The researchers also found that different emoji can have different sentimental connotations, and that these connotations can vary across different cultures and languages.

The paper "Financial Sentiment Lexicon Analysis" is a research article that discusses the development and evaluation of a financial sentiment lexicon for sentiment analysis of financial news articles[2]. The study was conducted by researchers at the University of Lisbon and was published in the Journal of Information Science in 2018.

The goal of the study was to develop a financial sentiment lexicon that would allow for more accurate sentiment analysis of financial news articles. The researchers created the lexicon by manually annotating a large data set of financial news articles with sentiment labels and then using this data to identify relevant financial terms and their associated sentiment polarity.

The proposed financial sentiment lexicon was evaluated on a data-set of financial news articles and was found to be effective in accurately identifying sentiment polarity. The study also compared the proposed lexicon with other existing sentiment lexicons and found that it outperformed them in terms of accuracy and coverage.

The paper "Sentiment-aware Emoji Insertion via Sequence Tagging" is a research article that proposes a method for automatically inserting emoji into text to convey sentiment. The study was conducted by researchers at Carnegie Mellon University and was published in the Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.

The goal of the study was to develop a method for inserting emoji into text that accurately reflect the sentiment of the message[3]. The researchers proposed a novel approach based on sequence tagging, which involves identifying the sentiment of individual words and then selecting an appropriate emoji to insert into the text.

The proposed method was evaluated on a data set of Twitter messages and was found to be effective in accurately capturing the sentiment of the messages and selecting appropriate emoji to convey that sentiment. The study also compared the proposed method with other existing methods for sentiment-aware emoji insertion and found that it outperformed them in terms of accuracy and efficiency.

The paper "Emoji and Emoticon in Tweet Sentiment Classification" is a research article that investigates the impact of emoji and emoticons on sentiment classification of tweets[4]. The study was conducted by researchers at Beijing University of Posts and Telecommunications and was published in the journal Information Processing & Management in 2020.

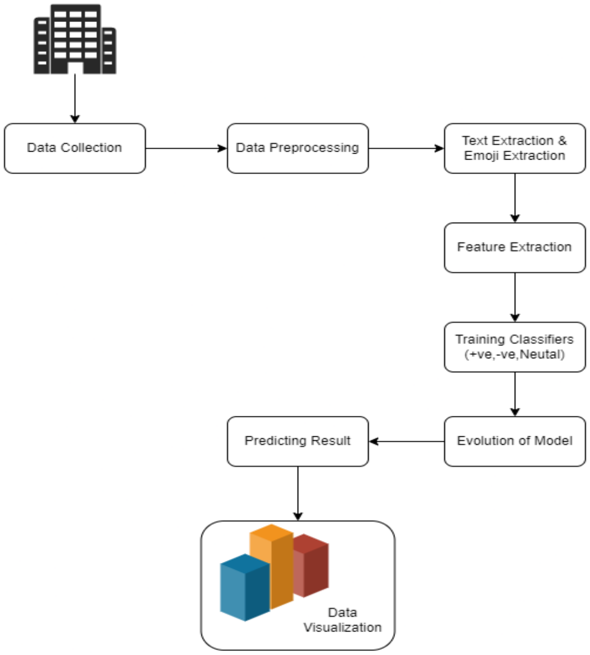
The goal of the study was to explore the effectiveness of using emoji and emoticons in sentiment classification of tweets, and to determine whether their inclusion could improve the accuracy of sentiment classification models. The researchers used a data set of tweets that included both textual content and emoji/emoticon data, and compared the performance of sentiment classification models that included or excluded this data.

The study found that including emoji and emoticon data in sentiment classification models can improve their accuracy, particularly in cases where the textual content alone is insufficient for determining sentiment. The researchers also found that different emoji and emoticons can have different sentimental connotations, and that these connotations can vary across different cultures and languages.

III. OBJECTIVE

1. Downloading product reviews on various websites featuring various products specifically amazon.com.
2. Analyze and categorize review data.
3. Analyze sentiment on data-set from document level (review level).
4. Categorization or classification of opinion into- • Positive • Negative • Neutral

IV. ARCHITECTURE



V. ALGORITHMS

Algorithm 1 – Lemmatization

Lemmatization is a textbook normalization fashion used in Natural Language Processing(NLP) that switches any kind of a word to its base root mode. Lemmatization is responsible for grouping different inflected forms of words into the root form, having the same meaning. Tagging systems, indexing, SEOs, information reclamation, and web hunt all use lemmatization to a vast extent. Lemmatization generally involves using a vocabulary and morphological analysis of words, removing inflectional consummations, and returning the wordbook form of a word(the lemma). The morphological analysis would need the birth of the correct lemma of every word. To simplify it, let’s just say that lemmatization is a verbal term refers to the act of grouping together words that have the same root or lemma but have different bows or derivations of meaning so they can be anatomized as one item. The process of lemmatization seeks to get relieve of inflectional suffixes and prefixes for the purpose of bringing out the word’s wordbook form.

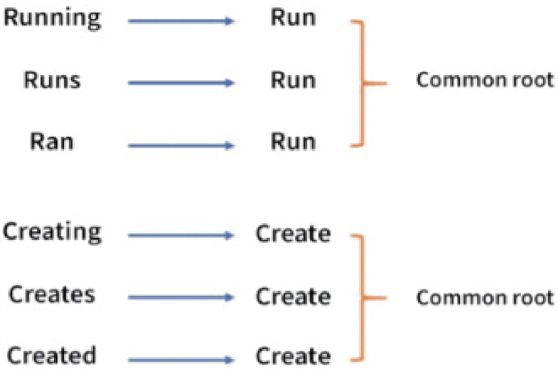


Fig.Lemmatization

Algorithm 2 – Support Vector Machine(SVM)

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Bracket as well as Retrogression problems. still, primarily, it's used for Bracket problems in Machine literacy. The thing of the SVM algorithm is to produce the stylish line or decision boundary that can insulate n- dimensional space into classes so that we can fluently put the new data point in the correct order in the future. This stylish decision boundary is called a hyper-plane. SVM chooses the extreme points/ vectors that help in creating the hyper-plane. These extreme cases are called as support vectors, and hence algorithm is nominated as Support Vector Ma spine. Consider the below illustration in which there are two different orders that are classified using a decision boundary or hyper-plane.

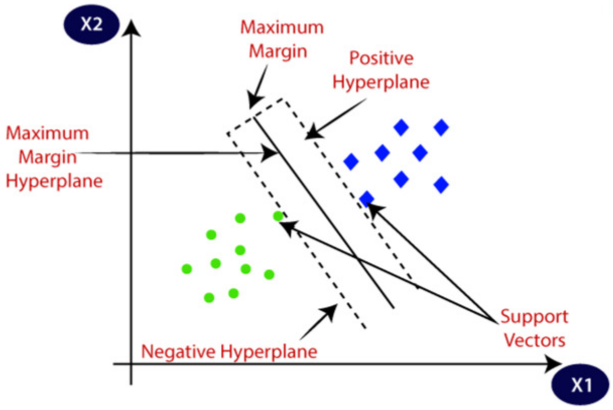
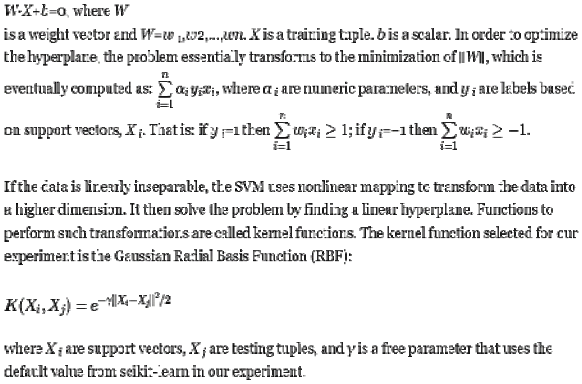


Fig.SVM

VI. Mathematical MODEL

Support vector machine (SVM) is a method for the classification of both linear and nonlinear data. If the data is linearly separable, the SVM searches for the linear optimal separating hyper-plane (the linear kernel), which is a decision boundary that separates data of one class from another. Mathematically, a separating hyper-plane can be written as:



$W{\cdot}X{+}b{=}o$, where W

is a weight vector and $W{=}w_1,w2,...,wn$. X is a training tuple. b is a scalar. In order to optimize the hyperplane, the problem essentially transforms to the minimization of $\|W\|$, which is eventually computed as: $\sum_{i=1}^{n} \alpha_i y_i x_i$, where α_i are numeric parameters, and y_i are labels based on support vectors, X_i. That is: if $y_i{=}1$ then $\sum_{i=1}^{n} w_i x_i \geq 1$; if $y_i{=}{-}1$ then $\sum_{i=1}^{n} w_i x_i \geq -1$.

If the data is linearly inseparable, the SVM uses nonlinear mapping to transform the data into a higher dimension. It then solve the problem by finding a linear hyperplane. Functions to perform such transformations are called kernel functions. The kernel function selected for our experiment is the Gaussian Radial Basis Function (RBF):

$$K(X_i, X_j) = e^{-\gamma\|X_i - X_j\|^2/2}$$

where X_i are support vectors, X_j are testing tuples, and γ is a free parameter that uses the default value from seikit-learn in our experiment.

VII. EXPECTED RESULT

Our system will produce a text output out which will describe the text and emoji nature whether it is Positive, Neutral or Negative .

VIII. CONCLUSIONS

Text and emoji classification is a field of study that analyzes people's opinions, attitudes, or emotions towards certain entities. We tackle a fundamental problem of Text and emoji classification. Online product reviews from Amazon.com are selected as data used for this study. A Text and emoji classification has been proposed along with detailed descriptions of each step. We will be using NLTK (Natural Language Toolkit) feature in python for further implementation sample review data. This will focus upon using in-built classifier models from NLTK package in python and compare their accuracy for a given data set.

ACKNOWLEDGMENT

We take this opportunity to thank our project guide Prof.R.K.Taware, Head of the Department Prof.S.A.Dabhade and Honorable Principal Dr.Mukane S.M. for their valuable guidance and for providing all the necessary facilities, which were indispensable in the completion of this project report. We are also thankful to all the staff members of the Department of Computer engineering of SVPM's College of Engineering, Malegaon(Bk) for their valuable time, support, comments, suggestions and persuasion. We would also like to thank the institute for providing the required facilities, Internet access and important books.

REFERENCES

[1] 2017 International Conference on Computational Science and Computational Intelligence "Sentiment Analysis of Tweets Including Emoji Data"Travis LeCompte and Jianhua Chen Division of Computer Science and Engineering Louisiana State University Baton Rouge, LA 70803-4020, USAe-mail: {tlecom3, cschen}@lsu.edu

[2] 2018 IEEE 12th International Conference on Semantic Computing (ICSC),"Financial Sentiment Lexicon Analysis", Sahar Sohangir; Nicholas Petty; Dingding Wang.

[3] Sentiment-aware Emoji Insertion via Sequence Tagging, Fuqiang Lin, Yiping Song, Xingkong Ma, Erxue Min, Bo LiuNational University of Defense Technology, China,2021.

[4] Information Technology International Seminar (ITIS) Surabaya, Indonesia, October 14-16, 2020. "Emoji and Emoticon in Tweet Sentiment Classification" Amalia Anjani, Eka Dyar Wahyuni Arifiyanti Department of Information Systems Universitas Pembangunan Nasional "Veteran" Jawa Timur Surabaya, Indonesia amalia_anjani.fik@upnjatim.ac.id

[5] 2018 IEEE 12th International Conference on Semantic Computing (ICSC),"Financial Sentiment Lexicon Analysis", Sahar Sohangir; Nicholas Petty; Dingding Wang.herjee A, Liu B, Glance N (2012) Spotting fake reviewer groups in consumer reviews In: Proceedings of the 21st, International Conference on World Wide Web, WWW '12, 191–200.. ACM, New York, NY, USA.

[6] S.P. Karunathilake; J.L.A.J. Shamal; R.G.H. Pemathilake; G.U. Ganegoda,"Feature Extraction from Online User Reviews", 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer).