International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 3 Issue 6, October 2019 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

# An Empirical Study on Comment Classification

**Shubham Derhgawen, Rajesh Tak, Himaja Gogineni, Subhasish Chatterjee**

Department of Information Technology, Dhole Patil College of Engineering, Pune, Maharashtra, India

**ABSTRACT**

Due to increasing technologies in the interactive web applications, there has been a lot of development in E commerce and online social networking activities. The comments or the post always plays a vital role in understanding of the attitude towards a particular topic, product of the online users. Most of the times these comments or posts help the other users to understand the scenario and to take the right decision on the web platform. Machine learning plays a vital role to understand and to estimate the accurate semantics of these posts and comments. Natural language processing is widely used for this, Most of the times natural language processing does not yield much expected results in the classification of these comments due to the complexity in the narration. These complexities generally arise either due to poor narration of the comments or highly sarcastic contents in the comments. So to overcome these problems this paper broadly studies all the past work on comment classification and try to find the new way of machine learning to get the highly classified labels of the comments.

***KEYWORDS:*** *Natural Language Processing, Machine learning, Comment classification, labels*

***How to cite this paper:*** Shubham Derhgawen | Rajesh Tak | Subhasish Chatterjee "An Empirical Study on Comment Classification" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3 | Issue-6, October 2019, pp.332-335, URL: https://www.ijtsrd.com/papers/ijtsrd28053.pdf

IJTSRD28053



## I. INTRODUCTION

Communication in humans has evolved drastically since the early ages. Humans in the early stone age did not develop the extensive language that we are accustomed to today but relied heavily on sign language assisted with a number of different calls. These calls then slowly took the form of a language, with various different words having a different meaning.

The language was developed due to the need to convey the thoughts and opinions. The language also helped our ancestors to remain together in groups peacefully and also help coordinate various tasks between them. It was highly useful to hunter and gatherers which relied on language to hunt effectively as a team and also gather useful resources and avoid the harmful items.

The language was further refined through the ages and the vocabulary has been ever-increasing to accommodate the various emotions, opinions, and expertise. Language is very inherent to humans and highly critical to our survival. It helps convey various messages and also contribute to society as a whole.

Language has enabled humans to achieve a lot of advancements, this is through the accumulation of years and generations of research that is available in the form of books. The books were only possible due to the invention of language and have enabled humans to reach even further heights through the knowledge acquired and spread across the world.

There are a number of languages around the world, most of them are very similar to one another and almost all of them are composed of words which are in turn composed of individual letters. This is due to the way a human brain processes the information. Which is very different from a computer which can only organize or understand structured data.

The computer can work with very high efficiency on structured data such as tables in a database and spreadsheets. Which is very different from the way humans perceive information or language. Most of the information in the world consists of raw text in various languages which is unstructured. Therefore, one of the most demanding tasks is to translate this information or design a computer that can understand and perceive unstructured languages such as language and speech. This is a very challenging task to achieve which can be done with the help of Artificial Intelligence, which can emulate the various processes that happen in a human brain while it processes information.

There has been a lot of interest in a number of programmers since the advent of computers and computer languages to develop programs that are capable of understanding human language. As humans have been writing for thousands of years now, it would be highly helpful if the computer could understand the written text.

Computers still can't effectively understand the English language thoroughly like another human would be able to, the understanding is highly limited due to the complex

nature of the language and the various unspoken rules that have been ingrained in the human understanding of the language does not make it easier.

Therefore, NLP, which stands for Natural Language processing, is very crucial as a field where there has been ongoing research on this topic. Due to the fact that most of the English language is highly inconsistent and sometimes requires a lot of context and sub-context to understand a given statement at any point in time. This is problematic and not easy to solve as Natural Language Processing is a very broad field.

This is helped by a number of researchers performing a lot of research into the field of Natural Language Processing. Various updated libraries are now accessible and open for public use that can be utilized to achieve Natural Language Processing in any application. But due to the lack of logic and consistency in the English language, most of the time, machine learning gives a confusing output.

This can be traced to the fact that the extraction of meaning from a sentence written in English is extremely difficult. Therefore, for machine learning, the complicated problem needs to be broken down into smaller parts. These smaller chunks then can be individually processed by the machine learning algorithm by chaining, this process is called Tokenization and help achieve efficient machine learning models.

Sarcasm is one of the most difficult and almost impossible to detect by the traditional form of Natural Language Processing applications. Sarcasm detection is a very daunting task for a computer, due to the fact that not most humans can still detect sarcasm effectively. Sarcasm in a language is something that is used to express a bitter taunt or a gibe aimed at something or someone. It is highly challenging to ascertain the sarcasm in normal speech as it can seem to pass by undetected. Most of the sarcasm tends to be contradictory in nature and is usually of a polarizing nature, either objectively negative or objectively positive. This is usually very difficult to detect.

Another characteristic of the sarcastic comments that are very complicated is that most of the sarcastic statements refer to various different events at the same time. This taps into the general knowledge, logical reasoning and anaphoras resolution to fully grasp the underlying meaning of the statement.

Sarcasm analysis is one of the most essential concepts to develop an accurate model for the Natural Language processing and the eventual classification of the comments and limiting the amount of spam that occurs online, especially on the social media websites. Humans tend to be highly hateful and such poisonous acts need to be nipped in the bud and an automated system capable of achieving this is very necessary.

This paper dedicates section 2 for analysis of past work as literature survey and section 3 concludes the paper with feasible statement of the literature study.

## II. LITERATURE SURVEY

Siswanto [1] explains that motorsports such as the MotoGP attract a lot of audience in the electronic and print media. This is highly evident in the electronic media due to the large advancements in the cyberspace ushering the new digital era. This makes it easier for the people to access the information through social media, the website or the comments on certain trending topics on the internet. Therefore, the researchers in this paper utilize the comments on social media websites such as Twitter on the topic of MotoGP with the help of naïve Bayes theorem and Support Vector Machine to mine the text. The major drawback of this research is that the accuracy has been capped at 95.5% and cannot be further increased.

N. Chandra states that there has been increasing contributions to the online community such as social media websites and blogs. This increases the incidence of some anti-social elements in society trying to pollute people's brains by commenting. This is a very malicious activity and can even be highly racist towards certain communities. Therefore, to reduce the instances of this activity, the authors have applied machine learning algorithms to classify the anti-social comments based on the K- Nearest Neighbors algorithm [2]. The major drawback in this technique is the lack of conclusive experimental evidence of the superiority of this technique.

A. Ikeda elaborates that there has been the addition of an indispensable commodity in the common people's lives, videos. Videos form an irreplaceable component of a person's life where millions watch a really high number of videos, most of these videos have a discussion section dedicated towards active public comments. These particular comments provide a lot of information and insight into the public opinion as anyone can post a comment contributing to the discussion which is highly useful for providing advertisements. Therefore, the authors have annotated the comments on the Nico Douga website which posts video streams [3]. The annotations are successfully used to classify the comments. The drawback in this study is that the authors have done annotation only on the referred contents, which has room for further improvement.

F. Prabowo introduces the most popular social media website around the world and especially in Indonesia, Instagram. The Instagram social media caters to a large number of users and is also home to a lot of sellers who are actively displaying their content and selling items through social media. All of the posts on the website can be commented on and most shop owners communicate through the comment section with their customers. Therefore, the researchers have utilized a statistical approach to classifying the comments [4]. The authors have deployed SVM (Support Vector Machine) and CNN (Convolutional Neural network) to achieve the classification which is based on feature extraction. The proposed model achieves an accuracy of 84% which is one of the major drawbacks of this approach.

M. Takeda proposes an innovative technique for the classification of comments posted on various web services and social media websites such as twitter, amazon, etc. Most of the time the research regarding the text classification is done the most basic approach has been to use the Bag of Words technique. The researchers in this paper have not utilized this but instead, have used the tree kernels to classify the tweets and other comments [5]. The authors have also deviated from the research to create a video

retrieval service for tourism-related content and have left a lot of room for improvement in the classification model.

M. Ibrahim informs that there has been a steady increase in the amount of harassment online and cases of cyberbullying. This is due to the large advancements in technology and access to a social media and internet that has led to a rise in these cases which is a cause of concern to a lot of online communities and public forums [6]. Therefore, the authors present an innovative technique for the classification of comments using deep learning and data augmentation techniques. The researchers utilized CNNs (Convolutional Neural Networks), LSTM (Long Short-Term Memory) to achieve promising results. The researchers only classified the toxic comments on the website which is a very limited approach to this classification task.

W. Jitsakul takes a different approach towards the comment classification task, due to the fact that most reviews and comments can be tagged and reported as spam, inappropriate. Positive, negative, etc. this feedback mechanism is utilized with the combination of a text classifier. The authors in this study have attempted to classify the comments by the customers on the basis of positive or negative remarks. The system has been demonstrated with the helo of experiments conducted on a dataset and achieves an average classification accuracy of 80% which can be further improved. [7]

M. Andriansyah [8] explains that social media has been the greatest source of most cyberbullying cases being reported. This is mostly done on the discussion forums and comments on various posts that have been posted on social media websites. To ameliorate this effect, the authors have designed a methodology for the classification on comments on a social media website such as Instagram with the help of SVM (Support Vector Machine). The researchers have applied this technique on the Indonesian Instagram dataset and achieved satisfactory results. But the major drawback of this research is that average classification accuracy achieved is 70% which is very less.

J. Savigny elaborates that there has been an increase in video streaming websites on the internet, especially YouTube which has been very popular around the globe with an extremely large user base. All of the videos posted on the website have a discussion space where the users can post their opinions in the form of comments. The researchers in this study aimed to perform the emotion classification of the comments based on labels such as fear, disgust, surprise, angry, sad and happy [9]. The authors utilized Word embedding to achieve their classification. The experimental results on a YouTube database yielded an accuracy of 76% which is very low for a classification technique.

T. Peng introduces phishing attacks which are one of the most dangerous and common attacks that are being utilized by attackers nowadays. Phishing attacks are also one of the least defended security attacks in this day and age. Therefore, to increase the security and provide a solution to this problem, the authors present an innovative technique based on NLP (Natural Language Processing) to analyze and detect the signs of a phishing attack [10]. The researchers perform a semantic analysis of the text on a website to determine if it has been created with malicious intent to steal consumer data. The major drawback in this technique is that it has not been tested adequately to perform its performance analysis.

H. Shen states that there has been an increase in the variety as well as the number of logs being produced every day. As there has been an increase in the production of logs of various sizes and types, the processing accuracy and speed of various technique utilized for the purpose of analysis of logs has been decreasing. Therefore, the management and querying of the logs have led to a significant increase in the cost. Therefore, the authors in this paper present a unique technique for the analysis of based on log layering and NLP (Natural Language Processing). The experimental analysis on a database of logs revealed that the proposed technique has improved the system by 40%. The major drawback is that the computational complexity has been reduced with the help of NLP but it has not been fully utilized to its potential. [11]

K. Sintoris explains that business process models are a highly useful resource and vital for the organization for its growth and stability. Therefore, the authors in this paper propose a technique for the extraction of Business Process model with the implementation of NLP (Natural Language Processing) techniques [12]. The researchers applied the Natural Language Processing to generate various business process models automatically based on the existing documentation. This study is a good example of how NLP can be used to extract a lot of information from a given text efficiently. The major drawback in this paper is that this technique has not been effectively experimented for performance evaluation.

P. Gupta elaborates the difficulties that arise when managing modern databases as most of them contain a large amount of structured data which is highly problematic to analyze. But it is essential to process all that information to gain valuable insight into the process and the data. Thus, to eliminate this problem, the authors in this paper have proposed an Intelligent Querying system that deploys a Natural Language Processing technique to achieve highly accurate processing of the Queries and in turn the database [13]. The major drawback in this technique is the high time and space complexities of this system.

Z. Zong introduces Natural Language Processing as a system for a machine-based automatic translation. The authors state that there has not been an efficient and accurate automatic translation system that can translate the given text accurately into another language. The researchers believe that this gap in the system can be bridged with the help of the implementation of the Natural Language Processing technique. As NLP is one of the most powerful approaches to understand the semantics of the human language, it would be highly helpful in achieving high levels of accuracy if applied in translation. The major drawback in this paper is that it is a statistical study and not a complete implementation of this technique. [14]

P. Rani [15] presents an innovative system for a voice-controlled home automation system, which is reliant on IoT (Internet of Things), NLP (Natural Language Processing) and AI (Artificial Intelligence) to achieve its goals. The authors have utilized the Internet of Things platform to connect most of the electronic appliances in the home. The Natural

Language Processing is used to interpret and understand the voice commands with the help of Artificial intelligence to achieve Home Automation. The major drawback in this technique is that the NLP application is highly limited and can be improved further.

## III. CONCLUSION

Classification of the comments which are posted on the web is always a tough task. This is due to the complexities lies in the narration of the comments by the different users, which unable the artificial intelligence system to take much more effort to create the classification labels. This classification of the comments broadly helps to understand and undermine the possible usage of the comments to the system or the other users. This paper studies most of the traditional and new arenas of the comment classification system to unleash the backdrop of the major problems. As a solution for this , this paper extensively relays on the machine learning concepts to handle the sarcasm and complex narration of the comments to classify them according to different labels like negative, positive, neutral and many more.

## REFERENCES


[1] Siswanto et al, "Classification Analysis of MotoGP Comments on Media Social Twitter Using Algorithm Support Vector Machine and Naive Bayes", International Conference on Applied Information Technology and Innovation, 2018.

[2] N. Chandra et al, "Anti-Social Comment Classification based on kNN Algorithm", International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2017.

[3] A. Ikeda et al, "Classification of Comments on Nico NicoDouga for Annotation Based on Referred Contents", 18th International Conference on Network-Based Information Systems, 2015.

[4] F. Prabowo and A. Purwarianti, "Instagram Online Shop's Comment Classification using Statistical Approach", 2nd International Conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2017.

[5] M. Takeda et al, "Classification of Comments by Tree Kernels Using the Hierarchy of Wikipedia for Tree Structures", 5th IIAI International Congress on Advanced Applied Informatics, 2016.

[6] M. Ibrahim et al, "Imbalanced Toxic Comments Classification using Data Augmentation and Deep Learning", 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018.

[7] W. Jitsakul et al, "Enhancing Comment Feedback Classification using Text Classifiers with Word Centrality Measures", 2nd International Conference on Information Technology (INCIT), 2017.

[8] M. Andriansyah et al, "Cyberbullying Comment Classification on Indonesian Selebgram Using Support Vector Machine Method", Second International Conference on Informatics and Computing (ICIC), 2017.

[9] J. Savigny and A. Purwarianti, "Emotion Classification on Youtube Comments using Word Embedding", International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA), 2017.

[10] T. Peng, I. Harris, and Y. Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning", 12th IEEE International Conference on Semantic Computing, 2018.

[11] H. Shen et al, "Log Layering Based on Natural Language Processing", International Conference on Advanced Communications Technology (ICACT), 2019.

[12] K. Sintoris and K. Vergidis, "Extracting Business Process Models using Natural Language Processing (NLP) Techniques", IEEE 19th Conference on Business Informatics, 2017.

[13] P. Gupta et al, "IQS- Intelligent Querying System using Natural Language Processing", International Conference on Electronics, Communication and Aerospace Technology, ICECA, 2017.

[14] Z. Zong and C. Hong, "On Application of Natural Language Processing in Machine Translation", 3rd International Conference on Mechanical, Control and Computer Engineering (ICMCCE), 2018.

[15] P. Rani et al, "Voice Controlled Home Automation System Using Natural Language Processing (Nlp) and Internet of Things (IoT)", Third International Conference on Science Technology Engineering & Management (ICONSTEM), 2017.