iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** V **Month of publication:** May 2020

DOI: http://doi.org/10.22214/ijraset.2020.5356

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue V May 2020- Available at www.ijraset.com

Sentiment Analysis: Statistical and Linguistic Approach

Shruti Nim[1], Aayushi Jain[2], Shweta Rajvanshi[3], Dr. Sweeta Bansal[4]
[1, 2, 3]*Department of Computer Science,* [4]*Assistant Professor, Inderprastha Engineering College, Ghaziabad, Uttar Pradesh*

Abstract: ***With the availability of the Internet and world moving, the trend of social media and blogging culture is growing all over the world. There is an abundance of data available online about every field. Along with this availability of data comes the need to extract useful information out of it. Since manual extraction of useful content from such vast amounts of data is very nearly impossible many automated techniques have been developed. Sentiment analysis is one such example. Sentiment analysis is a growing field at the intersection of linguistics and computer science, which attempts to automatically determine the sentiment, or positive/negative opinion, contained in text. This paper discusses Sentiment Analysis, its Applications, Challenges and Approaches.***

Keyword: Sentiment Analysis, Sentiment Classification, Naïve Bayes, Subjectivity Analysis, Linguistic Approach, Summarization.

I. INTRODUCTION

Sentiment analysis can be defined as a process of determining the effusive tone behind a stream of texts, accustomed for perceiving the opinions, attitudes and emotions expressed within a text. It can also be defined as "the process of computationally identifying and categorizing opinions expressed in a chunk of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral" [1]. Sentiment analysis or opinion mining deals with the computational treatment of opinions or sentiments and subjectivity in the text [2]. Now-a-days, a lot of users rely on online reviews. In April 2013, about ninety percent of customer's judgments depended on Online Reviews. In Computational Linguistics, the focus is on opinions rather than on sentiments, feelings or emotions. The terms 'opinion' and 'sentiment' are often freely substitutable as both specify the equivalent area of study.

II. APPLICATIONS

As social media has given people a chance to become more vocal about their opinions on products and services, many organisations have started to manage their supply chains to improve flexibility and responsiveness to meet requests as well as to work towards full customer satisfaction. Different areas where sentimental analysis is being used are enlightened in this section.

A. Intent Analysis

Intent analysis steps are used for analyzing the user's intention behind a message and identifying whether it relates to an opinion, news, marketing, complaint, suggestion, appreciation or query.

1) *Sentiment Analysis in Business Intelligence Buildup:* Having insights-rich information eliminates the guesswork and execution of timely decisions. With the sentiment data about your established and therefore the new products, it's easier to estimate your customer retention rate. Supported the reviews generated through sentiment analysis in business, you'll always suit this market situation and satisfy your customers during a better way. Overall, you'll make immediate decisions with automated insights. Business intelligence is all about staying dynamic throughout. Having the feelings data gives you that liberty. If you develop a giant idea, you'll test it before bringing life thereto. This can be referred to as concept testing. Whether it's a brand new product, campaign or a brand new logo, just put it to concept testing and analyze the feelings attached thereto.
2) *Voice of Customer:* Sentimental Analysis can analyze complex sentiments: "The hotel had a quick check-in, but room service was awfully slow." NetOwl understands that there are two sentiments contained here and returns a positive sentiment for the check-in but negative for the service. This is critical for getting down to the fine-grained analytical level necessary for a corporation to understand where it's making the grade and where it's failing.
3) *Governance:* Government's plans and schemes have always been a hot topic and brought about strong reactions from people. Topics can range from current bill passed to opinions about presidential election participants. Sentiment analysis helps to sort these views and perspectives to indicate the collective response of the people. Through Sentiment Analysis, the government keeps track of how people perceived its latest schemes. Political parties use people's reaction and opinions about them to style their strategies to woo people.

4) *Views and Perspectives:* Humans are social animals. Social Media and blogging culture has only catered this need. The growth in user-created media and information has led to the rise of social media and web, creating a significant publicly available data source. Whenever we feel strongly about something we tend to write it over our blogs or social media accounts. These views and perspectives, if categorized, can act as a major source of information to organizations. Different Social Media platforms monitor these views with the help of Sentiment Analysis/ Opinion Mining. "Trending" on Facebook and "Popular in your network" on Twitter tell users hot/streaming topics all over the world.

III. CHALLENGES OF SENTIMENT ANALYSIS

Sentiment Analysis is not just a social analytic tool but an arousing area of study. The competence to grasp exaggeration, sarcasm, positive feelings, or negative feelings has been troublesome, for machines that lack feelings. Algorithms have not been able to estimate with more than 60% accuracy the feelings which are conveyed by people. Nevertheless with so many shortcomings this is one area which is growing at great pace within many industries some of them are listed:

A. Grouping Synonyms

Synonyms as we know are the words that have the same meaning. We may have the situation where people use synonyms to tell the same emotion and opinions, hence it is necessary for the accurate classification of their sentiments.

B. Anaphora Resolution

Anaphora can be defined as a linguistic relation between two textual entities which is determined when a textual entity (the anaphor) refers to another entity of the text which usually occurs before it (the antecedent). The process of determining the antecedent of an anaphor is referred to as anaphora resolution.

C. Parsing

Basically it helps in identifying between what is the subject and object part of the sentence or which ones are verb/adjective.

D. Conference Resolution

It helps in identifying what a noun or a pronoun is referring to. It helps in improving the exactness of opinion mining.

E. Fake Reviews

It refers to activities like writing fake reviews (also called shilling) that try to mislead readers by giving false positive opinions to some target entities in order to promote that entity and/or by giving false negative opinions to some other entities in order to damage their reputations.

IV. CLASSIFICATION OF DATA AND DIFFERENT EXTRACTION TECHNIQUES

A. Classification

The data extraction first requires classification of it. The requirement to classify data goes parallely with extraction. How data is going to be extracted depends hugely on how we have classified it. The reason behind classification of data is that the various problems can be formulated easily by applying classification/ ranking/ regression. Two ways to classify data are polarity and subjectivity:

1) *Sentiment Polarity:* Tagging of an opinionated document as either positive or negative point or feeling is defined as sentiment polarity classification or polarity classification. This binary classification task is commonly called sentiment classification. An important objective of sentiment analysis is identifying the polarity of any given text at any level—whether the expressed opinion in a document is positive, negative, or neutral. There may be cases where the sentiments may not be strictly opinionated, for example classifying a news article as good or bad news is based on its subjectivity while for just a piece of news, it doesn't.

It is therefore useful to point out that:

a) The sentiment polarity of opinionated texts like "this dress is fabulous," objective information such as "impactful documentary" help to determine the overall sentiment.

b) The process of determining whether a bit of objective data is good or bad is still not easy as classifying it into one of several topic-based classes and hence is commonly faced by the challenges of sentiment analysis.

c) There's a very little difference between subjective and objective information. Is the documentary impactful?

The issue often faced with polarity is the analysis of comparative sentences. For example "This purse looks better." For author's evaluation against multiple view ratings can be used.

2) *Subjectivity Classification and Detection:* Textual information can be classified as: factual and opinionated information. Facts are the statement of actuality or occurrence. A fact is based on direct evidence, actual experience, or observation. On the other hand Opinions are statements of belief or feeling. It shows one's feelings about a subject. Solid opinions, while based on facts, are someone's views on a subject and not facts themselves. Subjectivity detection tends to acknowledge whether the given text expresses opinions or reports facts. Although sentiment classification and subjectivity detection are closely associated with one another, it's been reported that the separation of subjective and objective instances from text is a much tougher task than sentiment classification.

B. Feature

A plenty of work discussing feature selection for machine learning approaches, and learning techniques pertaining to extraction of information and categorization of text has been done [1]. Based on it, feature selection can be divided in two broader parts:

1) *Linguistic Approach:* This technique consists of clarification with the help of some set of rules and vocabularies making it important that a system be made with lexicons consisting of words and value of polarity [2].

a) *Term Presence v/s Frequency:* Frequencies and terms have always had a significant importance in text analysis since the beginning of time, many corporates use this technique to generate a WordNet where font size varies in accordance with the term frequency. Larger the font, more is the word frequency [1]. Another way is detection of the presence of a term rather than its frequency. The classification of terms can be done to detect polarity in a given text. One can assign binary 0 and 1 in the absence and presence of a term respectively. Frequency detection of a term may or may not help in sentiment analysis as the number of times a term is occurring will help in getting it emphasized but may not tell much about sentiment of the text [1]. This can be used while detecting subjectivity from a number of documents present. Presence of terms like "great", "nice" and "amazing" can help analyzers in subjectivity detection of a document to a large extent.

b) *Position of a Token:* Feature vectors are employed with the information about the position of a token as its position can play a very vital role in determining the tokens impact on subjectivity of unit for instance: a text in the middle has more impact than a text at the end [3,6] .The distance between different tokens can help to determine the variation of opinion.

c) *Part Of Speech (POS):* Parts of Speech like adjectives and adverbs relay a strong sense of opinion in a text hence largely help in detection of subjectivity. Many data-driven predictions have revolved around detection of adjectives in a document. There is a high correlation between the presence of adjectives and subjectivity of a sentence [1]. Some models take use of similar words to group together in order to determine polarity like good, amazing or nice are grouped together.

d) *Presence of Negation*

Another construction that changes the polarity of a text is Negation [4].But it can be a little challenging.

i) This show is good

ii) This show is not good may look similar but have opposite meaning.

iii) Never am I going to say that I hate horror comics

It shows a positive overall impact but usage of the word "never" and "hate" makes it hard to detect. ● One way of this is dealing with this is designing systems based on following algorithm:

(Negation + positive word) in a sentence →Increment negative sentiment value; (Negation + negative word) → Increment positive sentiment value.

e) *Irony, Sarcasm and the Likes:* Designing systems which can detect the presence of irony in a sentence can be a source of major headache. Mostly, irony plays the role of polarity reverser. Some researchers solely focus on algorithms for the detection, while some researchers try to identify the linguistic features for automatic detection of irony. Theoretically, it is assumed that irony suggests opposite or different meanings from what is said. Boundaries among irony and other figurative devices (sarcasm, satire or humor) can be misleading at times. One needs to keep in mind that different classifiers can give varying results while dealing with figurative languages. Though there is no standard definition of irony and sarcasm, developing classifiers for their detection is possible as the fact prevails that humans are able to detect irony easily [7].

2) *Statistical Tool:* This section consists of statistical models which is used by many data mining algorithms. For example- Naive Bayes Model. Some Twitter data mining tools are built using this model.

a) *Naive Bayes Model:* This model is useful for very large datasets. Despite its simplicity, the Naive Bayesian classifier performs more sophisticated classification methods.Naive Bayes classifier is an effective technique for text classification. While operating, Naive Bayes assumes a stochastic model of document generation.

We have used the assumptions that documents are generated according to a multinomial event model (McCallum and Nigam, 1998). Thus, a document is represented as a vector di = (xi1, ……., xi|v|) of word counts where V is the size of the vocabulary (vol.) Here vol = {w1, w2, …….., wi|v|}. Each xi t € {0, 1, 2, …….} indicates how often wt is present in a certain document Di.

Let us take model parameters p(wt|cj) and class prior probabilities p(cj) and assuming independence of the words, the most likely class for a document di is computed as:

$C^*(d_i)=\text{argmax}_j p(c_j)p(d|c_i)=\text{argmax}_j p(c_j)\ \pi_{t=1}^{|v|} p(w_t|c_j)^{n(w_t,d_i)}$

Where $n(w_t, d_i)$ is the number of occurrences of w_t in d_i . $p(w_t|c_j)$ and $p(c_j)$ are estimated from training documents with known categories.

3) *Summarization:* Summarization is done after the data is extracted.

Following are the ways to summarize the data:

a) *Single-Document Opinion-Oriented*

i) *Summarization:* Single document summary systems will generate a summary based on a single source document. Single document can be composed of some subdocuments with multiple paragraphs. The described content in each of these subdocuments emphasis on different aspects all surrounding the same topic. Generally single document is composed of different side information and the different side information is related to the local topic only

Graph-based summaries are very suitable where the information is a set of separate entities related to one another.

b) *Multi-Document Opinion-Oriented Summarization*

ii) *Textual Summaries:* Leveraging existing topic-based technologies is one of the techniques upon which systems have been developed. We can either use pre-existing algorithms to modify input or we can modify those pre-existing algorithms. It may happen that users may demand to view all the opinions or views available out. So, we can do textual summarization without using the topic-based techniques.

iii) *Non-textual Summaries:* Calculation of net polarity of reviews (negative/positive) is gathered from various documents. Summary statistics can be represented in the form of mean/average-centric forms using the thermometer-type.

C. Issues

Although Summarization has been a major source of problems to summarizers in natural language processing, especially in case of multi-document opinion oriented summarization.

Some of them are as follows:

1) *Authentication of Reviewer:* Authenticity of a document is introspective on account of the reviewer.
2) *Conflicting Issues:* It may happen that a single user can post conflicting reviews in a single document. This is a major source of problem during summarization
3) *Middle Rating Ambiguities:* Middle ratings may mean so-so or ratings of all features combined into a middle rating.
4) *Individual Tendencies:* Some reviewers have a habit of giving low review which in turn differ from their original views. On the contrary, others tend to be negative.
5) *Semantic Context:* Sometimes, Analyzers face various issues while deciding that number of passages refer to the same semantic context.

V. CONCLUSION

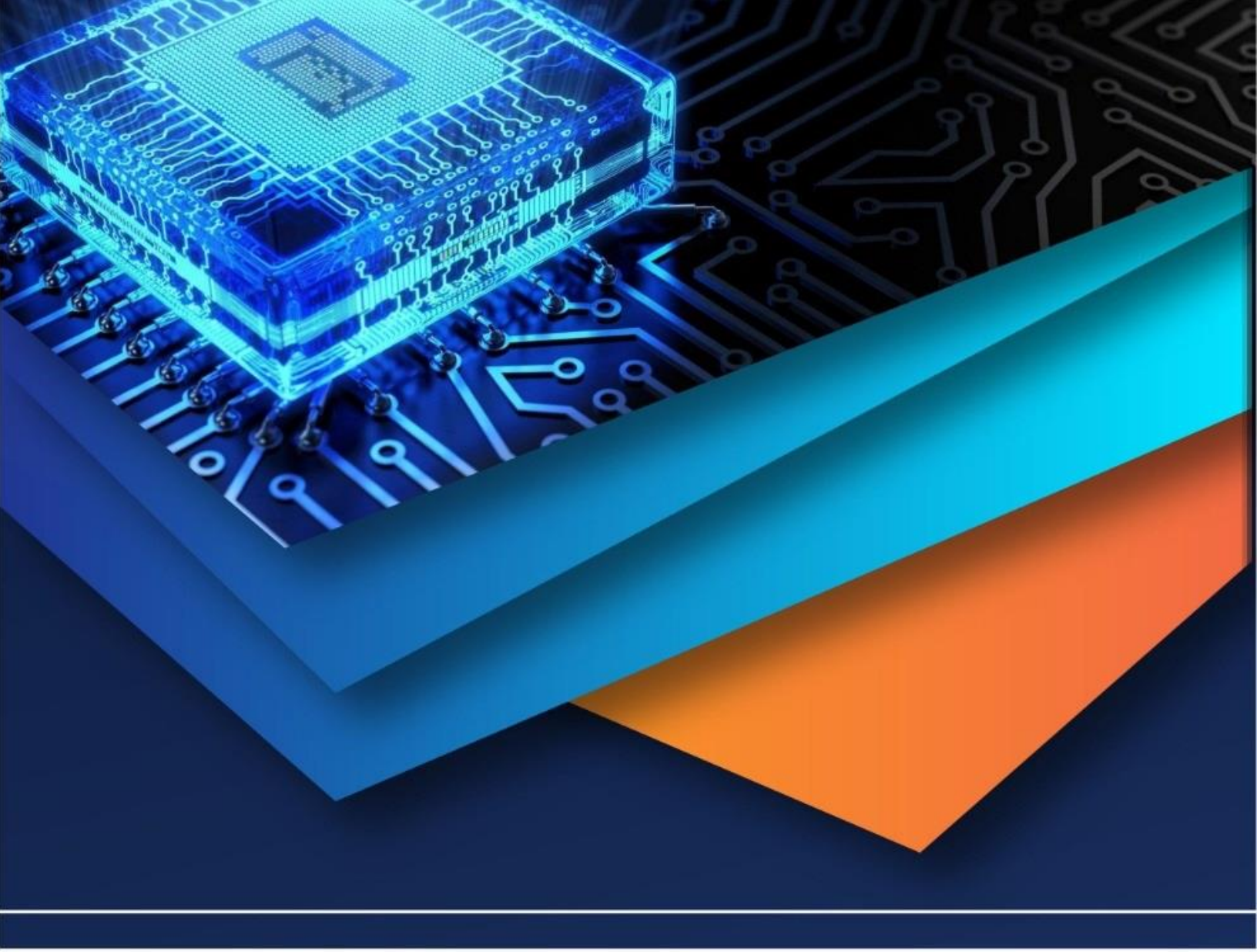
In this paper, we have covered various application challenges and the basic approaches associated with sentiment analysis with a special attention to Naïve Bayes method. Though not all aspects have been covered, the target of this paper was to highlight the major issues and the techniques related to sentiment analysis. We have also mentioned many challenges that researchers come across when doing Sentiment Analysis and summarization. With a large number of people getting interested in this area, future researchers will be able to tackle these issues. In addition to improvement in the current applications we will be able to use sentiment analysis in more and more applications.

VI. ACKNOWLEDGMENT

The authors of this paper are duly grateful to Dr. Sweeta Bansal, Assistant Professor, Computer Science Department, Inderprastha Engineering College for her help and support of this paper. Authors are also thankful to all other editors/reviewers and many anonymous reviewers for their comments. The paper is supported by the Department of Computer Science at Inderprastha Engineering College.

REFERENCES

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval 2(1-2), 2008, pp. 1–135.
[2] Preeti Routray, Chinmaya Kumar Swain, Smita Prava Mishra, "A Survey on Sentiment Analysis" International Journal of Computer Applications (0975 – 8887) Volume 76 – No.10, August 2013, pp. 1-6.
[3] S.-M. Kim and E. Hovy, "Automatic identification of pro and con reasons in online reviews," in Proceedings of the COLING/ACL Main Conference Poster Sessions, pp. 483– 490, 2006.
[4] Vaidehi Shah, Prof. Purvi Rekh "A Survey: Importance of Negation in Sentiment Analysis," International Journal of Emerging Technology and Advanced Engineering Volume 4, Issue 3, March 2014.
[5] Cristina Bosco, Viviana Patti, and Andrea Bolioli, "Developing Corpora for Sentiment Analysis: The Case of Irony and Senti–TUT (Extended Abstract)," Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015), pp. 4159, 2015.
[6] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 79–86, 2002.
[7] Cristina Bosco, Viviana Patti, and Andrea Bolioli, "Developing Corpora for Sentiment Analysis: The Case of Irony and Senti–TUT (Extended Abstract)," Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015), pp. 4159, 2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY