International Journal of Recent Development in Computer Technology & Software Applications
Vol. 5, Issue 1 – 2021
ISSN: 2581-6276

A Survey on Sentiment Analysis using Machine Learning

Shubham Jain[1]

[1]*Information Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu-603203.*

Abstract

In this review, various machine learning methods are used for opinion analysis. For the most part, did feeling examination by using AI classifiers like SVM (support vector machine), Random Forest, Naïve Bayes. In this, we see a few papers that help the new specialists establish an appropriate way to explore further. In this, there is a proposed strategy for the latest research program. Online media is the greatest medium to impart individuals' insights on various subjects. Feeling examination using AI strategies and with no human interference, machines will give individuals a precise opinion. Opinion study transforms text into positive, negative or impartial. Thus, any organization, establishment, or film commentator can take individuals' viewpoints and make further strides, as shown by that.

Keyword: Sentiment Analysis, Machine Learning, Naïve Bayes, AI, Web-based.

Introduction

Opinion mining is an AI used to dissect the texts for extremity from good to negative. Machine programmed figures out how to investigate the opinion of the human without the human information or interference. These days, web-based media is a piece of individuals' lives; individuals use web-based media to give their surveys over political fields, film audits, or advertising regions. There are various web-based media locales like Twitter, Facebook, Instagram and so on. They utilize these online media destinations as the medium to communicate their view on multiple points. In this way, opinion analysis investigates the text inputted by any individual from another country. Using the informational preparation index will explore the opinion of that specific text by knowing the feeling of that individual.

The opinion investigation application is exceptionally expansive and incredible, like Expedia Canada; Canadians use opinion analysis when they notice that individuals are giving negative remarks on the music used by their TV ads. Maybe rather than writing regrettable comments, Expedia realizes how to exploit that negative remark and air all-new genuine music in their channel.

A. Sentiment analysis Level

i. **Document-level:** Its investigation is utilized for the entire record. In this degree of arrangement, a report about an individual subject is incorporated. Clients have the attitude to think about two themes or two records. The managed and unaided AI procedures are used for the pattern of Document-level opinion analysis.

ii. **Sentence level:** Subjectivity arrangement is firmly identified with sentence-level opinion examination. The sentence-level opinion study discovers the expression like good, negative or unbiased from the given sentence. All the classifier from document level feeling research is utilized for sentence-level opinion analysis.

iii. **Aspect level:** The Aspect level opinion examination is used to discover feeling on the Aspect of those substances. "My vehicle has great dealing, but it is weighty" we should take this model. In this model, there is an assessment on a vehicle that the treatment of a feline is positive. However, the vehicle's weight is negative. The cutthroat statement is important for an Aspect level opinion investigation.

iv. **Phrase level:** In the expression where assessment words are discovered, their expression level arrangement is finished. These have benefits and burdens both because the use is there where the specific assessment on the element is there. Yet, in hindrance, there is logical extremity matter, so the outcome may not be precise.

v. **Feature Level:** Product include distinguishes as item credits; PRVdocument Analysing these provisions for recognizing feelings is called feature level opinion examination. The positive, negative or unbiased evaluation is distinguishing from split features.

Literature Survey

In this paper [1-2] Tweets are arranged into the positive or negative remarks utilizing AI algorithm, for example, Naïve Bayes, Random forest (RF), SVM, Unigram with Sentiwordnet unigram with Sentiwordnet including rejections are operating as the contribution to this paper. The author determined 3,000 one hundred 84 (3184) tweets using the tweeter API. 954 positive, 1318 negative, 145 stop words distinguished from 3184 tweets using. The author used opinion analysis features like Bag of words (BOW), TF-IDF, Unigram with Sentiwordnet, Unigram with Sentiwordnet, including invalidation words. The creator objects that every one of the classifiers with Unigram with Sentiwordnet and Unigram with Sentiwordnet, including contradicted words, shows higher accuracy than Bags of words (BOW) TF-IDF. Random forest with Unigram with Sentiwordnet, including negation words, get the most remarkable of 95.6% accuracy.

In this paper, [3] authors attempt to utilize an AI algorithm for Arabic client's input. They study two unique kinds of techniques which are casting a ballot and meta-classifier combination. They are gathering information utilizing Tweepy API [17]. There are numerous mocking and impartial tweets with positive and negative tweets. Collected a sum of 438,931 tweets from that 75,774 are positive and 75,774 negative. Eliminate all noisy information from the tweets like pictures, hashtags, retweets, and feelings; second, tokenization eliminates non-Arabic letters and

normalizes simple Arabic letters. Ten classifiers NB, ME, LR, RR, PA, MNB, SVM, SGD and Ada help BNB were utilized to extricate and find the extremity of given tweets. The highest accuracy accomplished by PA and RR is 99.96%. The most minimal accuracy achieved by Ada lift, LR and BNB are under 60%.

This paper [4] utilizes Amazon client survey information to discover the inspiration, pessimism, and nonpartisanship of clients' audits. They analyze two AI algorithms Naïve Bayes and the Support vector machine (SVM). The information is the client survey of the Amazon items. The investigation might be negative, positive or unbiased. Apriori algorithm is used to disengage the habitually used viewpoints from the information dataset. Sentiwordnet is used to compute energy, antagonism and neutral scores, and from that point onward, the classifier will apply. The correlation of the calculation depends on the exhibition can be determined by utilizing the Accuracy, Precision, Recall and F-1 Measure of every order. By the trial result, Naïve Bayes characterization is preferable accuracy over Support vector machine (SVM). The computation was finished by certain True positive (TP), False positive examples (FP), True negative examples (TN) and False-negative examples (FN).

In this paper, [5] Many natural email demonstrations are perhaps the greatest danger affecting the clients. The author joins both Sentimental investigation and character acknowledgement to examine the email content. They utilize two unique datasets to approve the proposed strategy. The first dataset is the first dataset (CSDMC 2010) and the second approval dataset (TREC 2007). CSDMC 2010 spam corpus:-This created 2949 messages to complete unique investigations. TREC 2007 public corpus:-There are 75419 messages in which 25220 are real 50199 spam messages. This technique was approved in two individual datasets working on the best precision in both the cases (from 99.15% to 99.24% and 98.98% to 99.18%). Further, this strategy is likewise utilizing for various approval like SMS and web-based media approval.

This paper [6] shows; During the pandemic of the COVID-19 entire world is languishing. Online media is a big stage to share your musings in any circumstance. The creator utilizes online media to examine individuals' responses to the present circumstance. The creator depicts the reality of how unreasonably individuals are acting in the present circumstance. It would be simpler for casualties to accumulate some organized data from online media. Two arrangements of datasets have been utilized in this paper. In dataset-1, there were 2,26,668 tweets utilized as the fundamental for dataset-2. They utilize the tweets which were retweeted most. To train the model, information has been sorted into train, approval and test sets. The accuracy of unigram, bigram and trigram was performed. The exactness of dataset 1 is 81%, and dataset 2 is 75% utilizing various classifiers. Overall, the author realized that online media isn't sufficiently valuable to help individuals.

In this paper, [7-11] creator look at the Alzheimer infection disgrace on Twitter utilizing AI strategy. AI method demonstrated disparagement communicated in 31150 Alzheimer sickness-related tweets gathered using tweeter API. In this 1% of the dataset used to prepare a classifier,

the tweet and the rest almost 100% of the dataset. In this paper, the creator discusses what online media outlets mean for disposition bearing in other advancement results. Taken out the retweet, different tweets which are not identified with Alzheimer were taken out, the watchwords "all", "Alzheimer", "dementia", "cognitive decline", "infirmity" characterized the example of examination. In conclusion, they eliminated the username which contains the subject name they destroyed. Two analysts manual coding and the result are 43.41% instructive, 23.79% joke, 21.22% allegorical, 19.29% association, 24.50% disparagement.

Issue Statement

It will be not difficult to contrast distinctive AI calculations to see which analysis performs better in various components. We can become more acquainted with the calculation's precision.

Proposed System

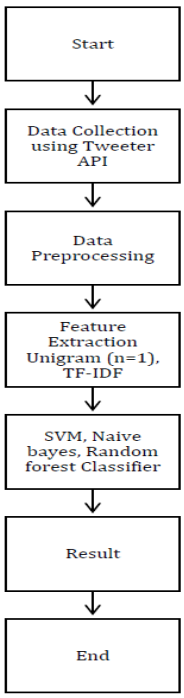

Figure 1.Flow chart of our Proposed Approach

i. **Information Collection:** Data assortment is the initial step of the feeling examination. There is an alternate assortment source like a blog, film survey, informal communication locales, item audit. Clients need to gather information utilizing Twitter API; for getting Twitter information, clients should make a Twitter account that gives buyer key, shopper mysterious, access token.

ii. **Information pre-processing:** In pre-processing of information, all the commotion from the dataset has been eliminated, as hashtags, URLs and designated names. Capitalized letters are changed over into lower case letters. Text tokenization has been done; tokenization is a cycle used to transform the text into a symbolic structure.

iii. **Element extraction:** Feature extraction is the main errand for grouping. All of the unimportant terms have been taken out from the dataset, similar to the word that doesn't communicate any feeling. For highlight extraction Unigram, term recurrence versus Inverse archive recurrence (TF-IDF).

iv. **Classifier:** The F-score is normally used to decide the exactness of a solitary classifier or think about the various classifiers. The equation of F-measure is given in the recipe. Multiple classifiers are utilized to group information like SVM (support vector machine), Naïve Bayes, Random timberland and so forth.

Result

In the outcome, various classifiers will be correlated to perceive how they get precision. Utilizing the provisions like Unigram (n=1) with Sentiwordnet, Bags of word, TF-IDF and so on. What's more, there will be diagrams likewise to show the yield and the working of the product.

Conclusion

Components like TF-IDF, Unigram and Bags of words by contrasting them and AI classifiers like SVM (Support vector machine), Random Forest and Naïve Bayes will not be difficult to show which element will get the best precision out of them. Likewise, it will give a yield in the chart, as it will be displayed in the table arrangement.

References

1. Soumya, S. and K. J. I. E. Pramod (2020). "Sentiment analysis of malayalam tweets using machine learning techniques."
2. Madhoushi, Z., A. R. Hamdan and S. Zainudin (2015). Sentiment analysis techniques in recent works. 2015 Science and Information Conference (SAI), IEEE.
3. Gamal, D., M. Alfonse, E.-S. M. El-Horbaty and A.-B. M. J. P. C. S. Salem (2019). "Implementation of Machine Learning Algorithms in Arabic Sentiment Analysis Using N-Gram Features." 154: 332-340.
4. Vanaja, S. and M. Belwal (2018). Aspect-level sentiment analysis on e-commerce data. 2018 International Conference on Inventive Research in Computing Applications (ICIRCA), IEEE.
5. Ezpeleta, E., I. Velez de Mendizabal, J. M. G. Hidalgo and U. J. L. J. o. t. I. Zurutuza (2020). "Novel email spam detection method using sentiment analysis and personality recognition." 28(1): 83-94.
6. Chakraborty, K., S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag and A. E. J. A. S. C. Hassanien (2020). "Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers-A study to show how popularity is affecting accuracy in social media." 97: 106754.
7. Narendra, B., K. U. Sai, G. Rajesh, K. Hemanth, M. C. Teja, K. D. J. I. J. o. I. S. Kumar and Applications (2016). "Sentiment analysis on movie reviews: a comparative study of machine learning algorithms and open source technologies." 8(8): 66.

8. Oscar, N., P. A. Fox, R. Croucher, R. Wernick, J. Keune, K. J. J. o. G. S. B. P. S. Hooker and S. Sciences (2017). "Machine learning, sentiment analysis, and tweets: an examination of Alzheimer's disease stigma on Twitter." 72(5): 742-751.
9. Li, Y. and H. Fleyeh (2018). Twitter sentiment analysis of new ikea stores using machine learning. 2018 International Conference on Computer and Applications (ICCA), IEEE.
10. Kolchyna, O., T. T. Souza, P. Treleaven and T. J. a. p. a. Aste (2015). "Twitter sentiment analysis: Lexicon method, machine learning method and their combination."
11. Shi, H.-X. and X.-J. Li (2011). A sentiment analysis model for hotel reviews based on supervised learning. 2011 International Conference on Machine Learning and Cybernetics, IEEE.

