**Abstract:-**

As of late, there has been a critical ascent in the internet business industry and all the more explicitly in individuals purchasing items on the web. There has been a ton of exploration being done on sorting out the purchasing behaviors of a client and all the more significantly the variables which decide if the client will purchase the item or not. In this investigation, we will explore on whether it is feasible to distinguish and foresee the buy aim of a client for an item and focus on that client towards the item with a customized ad or an arrangement. Further, we wish to foster a product that will assist the organizations with recognizing possible clients for their items by assessing their buy goal in quantifiable terms from their tweets and client profile information on twitter. In the wake of applying different content logical models to tweets information, we have discovered that it is to be sure conceivable to anticipate if a client have shown buy aim towards an item or not, and subsequent to doing some investigation we have discovered that individuals who had at first shown buy expectation towards the item have by and large likewise purchased the item.

**Introduction:-**

Several research studies have been conducted to examine the buying habits of online consumers. However, few have addressed the intention to buy products from customers. We want to develop an approach to machine learning which will identify potential product customers by estimating the purchase intention on tweet in a measurable way. We used a machine learning approach for text analysis because text analytics are inefficient, though they can be done manually. It will be much more rapid and efficient to find patterns and trends by using text mining and natural language processing algorithms.

Literary review:-

Several research studies have been conducted to examine the buying habits of online consumers. Only a few, however, have addressed the customers' intent to purchase things. Ramanand et al. (Ramanand, Bhavsar, and Pedanekar 2010) consider the challenge of finding ‘buy' wishes from product reviews in their study on identifying wishes from texts. These wishes can range from product suggestions to a desire to purchase a product. To detect these two types of wishes, they employed language principles. While rule-based approaches to recognizing wishes are successful, their breadth is insufficient, and they are difficult to broaden. The problem of detecting purchase intentions is similar to that of finding wishes in product reviews. We describe a machine learning strategy with generic features collected from tweets instead of a rule-based approach.

Previous research has shown that natural language processing (NLP) and named entity recognition (NER) may be applied to tweets (Li et al., 2012). (Liu et al., 2011). However, NER is challenging to apply to tweets since they frequently contain abbreviations, misspelt words, and grammatical errors. Finin et al. (2010) attempted to use crowdsourcing to annotate identified items in tweets. These techniques have also been utilised to apply sentiment analysis to tweets in other studies. Because product or movie reviews are either favourable or negative, the first research employed them. Wang et al. (2011) and Anta et al. (2013) looked at the sentiment of tweets that were filtered using a specific hashtag (keywords or phrases starting with the symbol that denote the main topic of a tweet). These studies only look at the sentiment of a tweet regarding a product that has been purchased by the author. We will, however, extract elements from tweets to determine whether the user has shown a desire to acquire the goods.

More recently, research articles such as Identifying Purchase Intentions by Extracting Information from Tweets (February 8, 2017, RADBOUD U NIVERSITY NIJMEGEN) and Tweetalyst: Using Twitter Data to Analyze Consumer Decision Process (The Berkeley Institute of Design) look into whether an artificial intelligence approach can predict (from existing user-created content on Twitter) if someone is a potential customer. Further reading of research papers such as The Impact of Social Network Marketing on Consumer Purchase Intention in Pakistan: Consumer Engagement as a Mediator (Asian Journal of Business and Accounting 10(1), 2017) gives us a better understanding of the impact of social network marketing on consumer purchase intention and how it is influenced by the mediating role of consumer engagement. According to the UGT theory (Uses and Gratification Theory).

The sentiment140 API (Sentiment140 allows you to discover the sentiment of a brand, product, or topic on Twitter), the TweetNLP library (a tokenizer, a part-of-speech tagger, hierarchical word clusters, and a dependency parser for tweets), unigrams, bigrams, and stemming are some common preprocessing techniques commonly used for twitter data. There are several dictionary-based approaches as well, such as using the textBlob library (TextBlob is a Python (2 and 3) module for textual data processing. It offers a unified API for common natural language processing (NLP) operations like part-of-speech tagging, noun phrase extraction, sentiment analysis, and more.

Linear Regression, Random Forest, Naive Bayes, and Support Vector Machine are some of the most often used machine learning methods for text analysis. We'll take a closer look at these models later.

**Proposed Approach:-**

In this section, we go over the specifics of our strategy to solving the challenge of detecting purchase intent. We'll start by going over our data gathering and annotation procedure. Then we'll go through how we go about preparing and modifying data in order to train text analysis models.

**Data Collection and annotation:-**

We had to make our own because there are no publicly available annotated Twitter tweets corpora for detecting buy intent. This was accomplished by crawling the webpage with a web crawler built by JohnBakerFish. We had gathered over 100,000 tweets, but because they were not labelled, we had to narrow it down to just 3200 tweets, which were chosen at random from the dataset and manually annotated using a set of criteria:

Criteria for Labelling of tweets

|  |  |  |
| --- | --- | --- |
|  | Tweet | Class |
| 1 | Comparing iphone x with other phone and telling other phone are better? | No PI |
| 2 | Talking about good features of iphone x? | PI |
| 3 | Talking about negative features of iphone x? | No PI |
| 4 | liked video on Youtube about iphone x? | PI |

Due to time constraints, we only used 3200 tweets from such a vast dataset. We defined Purchase Intention as an item that is related with action words such as (purchase, want, desire). Three people read each tweet, and the final class was determined by the most votes.

## Data preprocessing

### Data preprocessing techniques:

Next, we preprocessed the tweets using these techniques:

1. LOWERCASE: To achieve case uniformity, we began our foundation by changing our text to lower case.

2. REMOVE PUNC: We next used the punctuations and special characters removal feature to remove the lower case content. Unwanted special characters, spaces, tabs, and other elements may appear in text, but they have no bearing on text classification.

3. REMOVAL OF STOPWORDS: Text also contains unnecessary words that are part of the phrase and grammar but do not add to the meaning of the sentence. The terms "the," "a," "an," "in," and so on are among those mentioned above. As a result, these words are unnecessary and should be removed.

4. REMOVAL OF COMMON WORDS: There are also many repeating words that, due to their frequency, do not contribute to the meaning of the sentence. This might also be the consequence of a blunder, given the data we're looking at is unstructured and doesn't take formal sentence standards into account.

5. REMOVAL OF RARE WORDS: We also eliminated several uncommon words such as names, brand names (not iPhone X), and html tags. These are one-of-a-kind terms that don't add anything to the model's interpretation.

6. SPELLING CORRECTIONS: Social media data is riddled with grammatical errors. It is our responsibility to repair these errors and provide the proper word as input to our model.

7. STEMMING: After that, we went back to the basis of the terms. Stemming is done by cutting the word at the end or beginning and considering the common prefixes and suffixes found in that word. Porters Stemmer, which is included with NLTK, was used for our purposes.

8. LEMMATIZATION: After that, we lemmatized our text as well. This investigation is carried out in morphological sequence. A word is traced back to its lemma, and the output is the lemma.

We have roughly 1300 tweets for training data and the rest for testing after preprocessing the tweets.