JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Categorizing Posts Age wise Using Bayesian Classification Technique

1 Saneeya Maghrabi, 2 Dr Vilas Kamble
[1]ME Student, 2Professor
[1,2] Department of Computer Science & Engineering,
[1,2] PES College of Engineering, Aurangabad

Abstract

Online informal communities like twitter have a huge load of information. Nonetheless, consistently people don't give individual information, similar to age, sexual direction and other section data, yet the conviction examination uses such information to cultivate important applications in people's everyday schedules. Notwithstanding, there is at this point a mistake in this sort of examination, whether or not by the foreordained number of words contained in the word reference or in light of the fact that they don't consider the most different limits Can affect feelings in sentences; Therefore, more strong results will be gotten on the off chance that considering customer profile data and customer creating style. This investigation shows that perhaps the main boundary contained in the customer profile is the age pack, which shows that there is customary lead among customers of as old as, especially when these customers clarify. With a comparable topic Detailed assessment with 7000 sentences has been coordinated to sort out which components are critical, similar to the usage of emphasis, number of characters, sharing of media, various subjects, and which ones can ignore the age pack gathering. Various learning machine estimations have been pursued for the portrayal of adolescent and grown-up social events and the Word2Vec has the best show with precision up to 0.95 in the endorsement test. should additionally, to affirm the handiness of the proposed model for age bundle portrayal, it is executed in the Sentiment Metric (eSM) that has been improved.

Keywords Tweets,NLP, Twitter,Deep Learning.

Introduction

Users spend time browsing websites, e-commerce, reading sports, journalism and entertainment news, even expressing their opinions and feelings in the form of comments on social media on a variety of topics. These feedback can be analyzed to assess customer satisfaction as very useful information for service providers and product providers. Goldsmith and the faculty [1] explore the behavior of people who use the Internet for electronic commerce and emphasize the importance of evaluating customer satisfaction in this type of service with each other. There are many applications that are used that use confidence analysis, such as psychological illness detection [2], false profile detection [3] to prevent criminals from attracting new victims. [4] Predicting the success or failure of political candidates Measure the spread of disease and determine the level of crime in the city. [2] There are currently many concerns and efforts to analyze data from online social networks to anticipate information that may reflect various aspects of being True today [5].

Twitter social network due to data availability policy, there are many short sentences, tweets that can be collected and analyzed. However, short and informal sentences with a large number of languages [3] require improvement of some

parameters to analyze the data. Among them is the age that can directly influence the final sentences of the sentences [6], [7]. In this type of analysis, the general characteristics found in each period of life are considered. Especially those characteristics will be clearly different in adolescents and adults. It is important to note that in some social networks, the user's age cannot be used either in the social network itself or by the user for reasons according to the criteria. Therefore, determining the method of predicting the age of users is relevant to analyzing trust. There are few studies that take into account the influence of age and gender in a way that a person can express feelings on a blog. Blocks with precision up to 80.32% [9], [10]

De Jonge et al. [11] Check text message abbreviations for high school and college students, such as smileys, slang, lengths, text, and misspellings. Huffaker and the faculty [12] examine teenagers' language use in blogging. They concluded that the most used language are abbreviations and emoticons, Shapiro and the faculty. [13] Study how often teens write on social media. Each of these tasks uses specific parameters.

In this context, the main contribution of this work is to show that parameters such as the use of punctuation, including emotional icons, the number of characters in the text or the length of the sentence in jargon, the use of the Uniform Resource Locator (URL)) to share media information. , The number of people who follow, the number of followers, the total number of tweets published on social networks and the topics that are relevant to increase courage Expression and precision in the classification of age groups Some of these parameters are used in other applications such as jargon, emotion icons, and sentence lengths [14] - [16]; But they don't consider parameters such as score, URLs, people followed by users, followers, and the total number of tweets. Each of these parameters is determined after a qualitative analysis is performed manually and considers many sentences collected from Twitter. Furthermore, our research also considers which parameters can be ruled out when classifying age groups, such as the references used. @ Symbols, hashtags and shared messages Education levels are not considered in this work because they have been tested and have low precision. It should be noted that the context of various topics such as health, family, politics, work environment and other considerations are considered.

In addition, it is important to note that although the education has been carried out using social networks Twitter, but it can be extended to other social networks because the parameters are usually the same. In this research, to classify adolescents and young adults, different machine learning algorithms and neural networks have been tested. Convolutional (DCNN) to the best performance In addition, to determine the usefulness of the proposed model to classify age groups, it is implemented in the Sentiment Metric (eSM) that has been improved. [18] In performance audits, subjective tests are performed and the results are compared with the following confidence indicators SentimeterBR2 [19], [20], eSM, without considering the proposed eSM format with the proposed model and eSM that considers the actual age group. These results show the relevance to determine the parameters and benefits of the age group.

Literature Review

In this part, the primary investigation of the impact of specific boundaries in the order old enough gatherings has been noticed. There is likewise a conversation of certainty examination and AI calculations. In the opinion investigation, a few examinations have been refered to stress that the client's age information is a significant boundary in working on the productivity of estimating the power of sentiments. A. The connection between age gatherings and the idea of composing.

Brain science shows the distinction in the conduct of individuals in various ages. [21], [22] as a general rule, teenagers couldn't care less with regards to their protection. [23] and they post and distribute various data. A ton on informal communities online It can be viewed as a youngster as an individual of up to eighteen. That is the point at which they arrive at the period of greater part. Be that as it may, for social reasons, a few nations use somewhere in the range of 13

and 19 years. [21] Age data isn't accessible in some interpersonal organizations, like Twitter. In the wake of making sure that this data can really change the aftereffects of many investigates, research Something [3], [17], [24] attempting to foresee it.

One methodology utilized is to look for depictions in profiles with articulations "X years", "I have X years" or "I utilize X years" where X addresses the age of the client. Notwithstanding, it has been confirmed that on Twitter, the age profile in profile subtleties is anything but a typical propensity. [25] Therefore, these investigations won't give solid outcomes. It is normal among youthful clients of informal organizations to talk about different subjects that happen in their day to day routines, influencing their genuine world. [10] Topics, for example, school connections And companions frequently live in this age bunch. [12].

Mature clients identify with their own pictures. Then, at that point, they will be more cautious with the remarks they compose and the people who can peruse [26]; Therefore, it is feasible to observe sentences that have more good sentiments, not utilizing self-referring to, to utilize less disavowal. [8] Therefore, utilizing shoptalk turns out to be less [27]. Less with regards to yourself can be demonstrated when clients on the web. In adulthood, clients have more responsibilities for the duration of the day and youngsters invest more energy with online media than hours out of every day. Then, at that point, for teens, interpersonal organizations turned into a significant device to offer their viewpoints on the world. [13]

Notwithstanding personalities of customary character in grown-ups, like subjects of religion, belief system, legislative issues and work; Adults additionally utilize online media to offer viewpoints. Mature clients know about joining pictures, recordings or sharing connections of different pages that will fill the data that starts in tweets. [28] Between these two gatherings, Twitter clients younger than thirteen are not considered in this exploration on the grounds that numerous web-based informal organizations expect clients to Can be utilized for somewhere around thirteen years to have the option to apply Therefore, a gathering of teens comprising of clients matured 13 to 20 years, otherwise called youngsters and grown-ups, comprises of all clients matured 20 and over.

In this unique circumstance, it is normal to observe concentrates on that break down the assessments of online business sites and informal organizations by isolating the sensations of that remark. There are many investigations on certainty examination. Yet, most don't consider client profiles, for example, measurements Sentimeter-Br2 That depends on the word reference, which each word has a positive or negative worth of certainty This pointer considers n-grams, modifiers and no words to stop. Contrasts, values, certainty rely upon the verbal words where the previous action words have less certainty than the action words of right now. Sentimeter-Br2 That is situated on Sentimeter-Br [29].

The review depicted in the accompanying portrays the effect of client profile information on certainty examination fully intent on expanding the viability of the certainty pointers. 1) ANEW is a review that considers information related. Clients for certainty investigation [30] have concentrated on whether the presence or nonappearance of sex information will influence the final product. 2) SentiWordNet is a proportion of network. Clients are utilized to arrange polar certainty naturally means to aid the examination of web-based media. [31].

The creator dissects that it is so vital to consider the client profile and later changing the focal point of the review to cover different language examination. Clients, including sex, training level, geographic area, and age, eSM is a relationship of certainty markers that utilization vocabulary word reference, Sentimeter-Br2. With the altering factor dependent on the client's profile data, eSM arrangement of Fi sentences set by (1). Kindly note that there is a suspicion that that age data exists in the client profile.

Notwithstanding these work, others [33], [34], report that in investigating thorough opinion, it is vital to have markers for the assessment of the incongruity since this can invert the Convinced sentences In the two works, the age of the client can impact how to rank the sentences amusingly. Nonetheless, on Twitter, age data is private and not required.

Essentially, [12] the writer observed that youngsters act contrastingly in the internet based climate and it is feasible to notice specific qualities in the composing style, like themes that influence their reality. Setting up posts about subjects that don't allude to themselves and managing more certain data might be the attributes of mature clients [26]; While utilizing shoptalk is regularly found in sentences posted by teen clients. [27] Moreover, the need to connect the media that addresses the substance referenced is the attributes of grown-up clients. [28].

Consequently, this work is planned to bunch the qualities referenced before [12], [27], [28] and others proposed in this exploration, like the utilization of accentuation, truncations, images that express feelings and attributes of Writing style notwithstanding other client data, for example, history, tweets, number of devotees and the quantity of individuals the person follows.

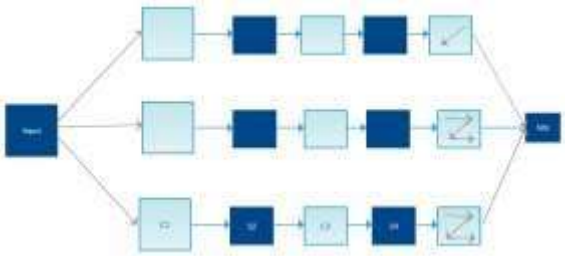
There are many AI models [35] that cover essential strategies, for example, direct relapse and tree models, including more refined techniques, for example, fake neural organizations or vector support. As a rule, AI isn't restricted to just a single information investigation. Yet, analyze many models and pick the model that has the best prescient precision Machine learning region, otherwise called design acknowledgment or information mining [36], includes the division of examples in huge informational collections. Frequently the objective is to foresee the right reaction variable, for instance, age bunches dependent on at least one prefixes, like composing style.

Subjective techniques are regularly not considered in research about certainty examination. [37] There is a need to channel or erase useless sentences that are considered the most vocal in the information distribution center by Twitter.] - [40] which has a great deal of data Many occasions in the beginning phases of information examination, it is important to indicate the primary qualities or examples of the example and this work is finished by the actual specialists. In this unique situation, research is directed on a lot of information and is separated and shrouded in a populace, paying little mind to explicit individuals. Then, at that point, considerably more close to home data, for example, "I miss you home" or "I'll begin lessening carbs." Do not share the client's very own data in the outcomes. Eventually, the objective is to attract designs that exist the way of articulation from each age bunch and not simply sometimes.

To have the option to work with a lot of information and accomplish the ideal grouping, the AI's calculation is utilized which can give high exactness results. [41] For the calculation example that Using choice trees (J48), vector support (SMO) or fake neural organizations, the utilization of deep Leaning expansions in numerous ways, which has happened as of late, for example, pictures [42], [43]. Also Strasbourg voice The deep inclining calculation permits the computational model that comprises of numerous handling layers to become familiar with the portrayal of information with various degrees of deliberation. As of late, deep learning strategies have been applied to message grouping. [44], [45] and the calculation has astounding outcomes for the grouping of text designs [46] - [49]

Deep learning is frequently deciphered as far as the widespread estimation hypothesis [50] or the likelihood derivation [51]; The guess hypothesis characterizes a class of all inclusive approximations, which alludes to the capacity of the neural organization of taking care of straightforwardly with a solitary, restricted size mysterious layer for roughly persistent capacities. Translation of likelihood comes from AI, including deduction, just as improvement ideas like preparing and testing identified with transformation and general attributes.

The DCNN can perform classification tasks, and it is composed of multiple layers, each one computes convolutional transforms [52]. Fig. 1 shows the topology of the DCNN, in which the C variables represent the convolution layers and the S variables represent the layers pool/sample; from C1 to S2 a convolution layer is present, from S2 to C3 a sub-sampling layer is present, from C3 to S4 exist another convolution layer, and from S4 to the output a fully connected Multilayer Perceptron (MLP) is represented.



Neural Network

Proposed Method

For classifying age groups that considers two phases, the data treatment extracted from social networks and the classification phase.

To obtain a more exact prediction of the age group from the tweets were extracted from the social network containing the written message and user profile information using deep learning techniques and neural networks.

To perform a detailed analysis with huge amounts of tweets to determine which characteristics are relevant, such as, the use of punctuation, number of characters, media sharing, topics, among others; and which ones can be disregarded for the age groups classification.

To use different machine learning algorithms and test for the classification of the teenager and adult age group, and the Word2Vec Model.

To draw patterns that exist in the manner of expressing themselves from each age group, and not just a few isolated cases.

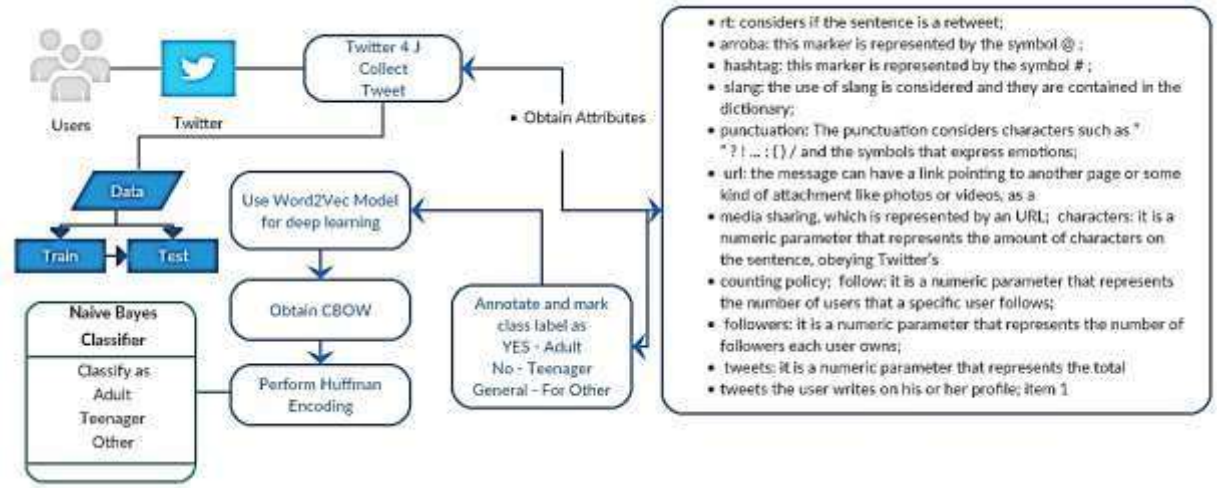
$$eSM(Fi) = \text{Sentimeter } Br2(Fi) * C * \exp(a1 * A1 + a2 * A2 + \ldots + an * An + g1 * M + g2 * F + e1 * G + e2 * nG + t1 * T2 + \ldots + tm * Tm) \quad \ldots\ldots(1)$$

Where:

- C is a scale constant, obtained by subjective tests
- a1... an are binary factors related to age groups A1...An are the weight factors of each age group, been considered four groups;
- g1 and g2 are binary factors related to the gender; M and F are the weight factors of gender, man or woman, respectively;
- e1 and e2 are binary factors related to educational level (higher education or not);
- G e nG are the weight factors of educational level, higher education or not, respectively.

In order to obtain a more exact prediction of the age group, some information extracted directly from the social network was considered and some parameters that were considered important during the tests for this research. Among them is the punctuation mark, which was considered to know if the user had written some type of punctuation in the message; commas and end-point are disregarded because they are more common in any type of sentence. In this entry the symbols that express emotions, the called emoticons, were also considered as being punctuation. This last parameter has already been used in gender detection.

The use of slang has been incremented with the predefined abbreviations in a dictionary, besides the spelling variations of the words. The entry that refers to the attached media, the URL is composed by the tweet messages that contain a link pointing to another page, or some kind of attachment such as photos or videos. We also considered if the message has the markers or symbols, “#” that highlight some topic, and “@” which is used to mention the name or nickname of another user.



Proposed System

Others entry parameters considered in this work are extracted directly from the user profile and they are part of his or her history on the social network. These are the number of people the user follows, the number of followers he or she owns, and the total number of tweets posted on his or her profile.

The following parameters are considered to predict age group:

- rt: considers if the sentence is a retweet;
- arroba: this marker is represented by the symbol @ ;
- hashtag: this marker is represented by the symbol # ;
- slang: the use of slang is considered and they are contained in the dictionary;
- punctuation: The punctuation considers characters such as " " ? ! ... : ( ) / and the symbols that express emotions;
- url: the message can have a link pointing to another page or some kind of attachment like photos or videos, as a media sharing, which is represented by an URL;
- characters: it is a numeric parameter that represents the amount of characters on the sentence, obeying Twitter's counting policy;
- follow: it is a numeric parameter that represents the number of users that a specific user follows;
- followers: it is a numeric parameter that represents the number of followers each user owns;
- tweets: it is a numeric parameter that represents the total tweets the user writes on his or her profile
- topic: the main topic of the sentence is considered;
- gender: the gender of the user who writes the message, which is represented by male and female genders;
- teenager: the parameter that represents the output of the machine learning algorithm, it is represented as teenager or no teenager (adult).

The parameters as rt, @, hashtag, slang, punctuation, URL and the definition whether the user is teenager or not are binary, because they have only the YES or NO response, if the answer is positive or negative, respectively. Also, the gender parameter is binary, in which the symbol F was assigned for woman, in the gender field, and M for man. The other entries: characters, follow, followers, tweets and topic are numeric parameters that represent the actual extracted.

Experimental Setup

Due to various programming languages & its compatibility issues with databases and utilization some lead to develop libraries for reusable patterns. In this paper, we explore the utilization of Twitter4J libraries which are reliable Twitter APIs and that can be integrated to any applications for data acquisition in any format. It is a cross-platform tool and can be used on several operating systems, with the latest versions of Java Runtime Environment. The utility can be used as it is without any customizations it has no dependencies to any other system on which it runs. The usage of Twitter4J is simple, as all you need to do is copy the JAR file to the preferred classpath and use it. Here we explore the method of using twitter4J libraries for data acquisition for data analytics. This work will help data scientist, data quality analyst and business users.

Twitter4J is an official Java library for the Twitter API. Twitter4J one can easily integrate any application with the Twitter service. Twitter4J has features such as, 100% runs on Java Platform version 5 or later, Android platform and Google App Engine ready, Zero dependency, No additional jars required, Built-in OAuthsupport, Out-of-the-box gzip support, 100% Twitter API 1.1 compatible. By adding twitter4j-core4.0.4.jar to any application class path. If you are familiar with Java language, looking into the JavaDoc should be the shortest way for you to get started twitter4j. Twitter interface is the one you may want to look at first.

The Twitter Volumes View indicates Tweet volumes related to tweets as histograms over the last 14 days prior. The histogram is shown for the prediction candidate and types of tweets selected for comparison, if Twitter data is available. By hovering over the bars the respective number of tweets is displayed and the Word count and Table View are updated to indicate most prominent terms and tweets of the selected day accordingly. Based on a specifically trained SVM classifier we can separate humans interest tweets (“Wanna watch #ironman ”) from cyborg tweets and buzz. These affections are consistently highlighted (red=humans, blue=cyborg/buzz) of overall volumes, tags and individual tweets in the three views. NLP techniques are based on machine learning and especially statistical learning which uses a general learning algorithm combined with a large sample, a corpus, of data to learn the rules . analysis has been handled as a Natural Language Processing denoted NLP, at many levels of granularity. Starting from being a document level classification task, it has been handled at the sentence level and more recently at the phrase level .NLP is a field in computer science which involves making computers derive meaning from human language and input as a way of interacting with the real world.

Unigram: Building the unigram model took special care because the Twitter language model is very different from other domains from past research. The unigram feature extractor addressed the following issues:

a. Tweets contain very casual language. For example, you can search "hungry" with a random number of u's in the middle of the word on http://search.twitter.com to understand this. Here is an example sampling: huuuungry: 17 results in the last day huuuuuuungry: 4 results in the last day huuuuuuuuuungry: 1 result in the last day besides showing that people are hungry, this emphasizes the casual nature of Twitter and the disregard for correct spelling. b. Usage of links. Users very often include links in their tweets. An equivalence class was created for all URLs. That is, a URL like "http://tinyurl.com/cvvg9a" was converted to the symbol "URL."

c. Usernames. Users often include usernames in their tweets, in order to address messages to particular users. A de facto standard is to include the @ symbol before the username (e.g. @alecmgo). An equivalence class was made for all words that started with the @ symbol. The query term affect the classification. 2. Bigrams d. Removing the query term. Query terms were stripped out from Tweets, to avoid having the reason we experimented with bigrams was we wanted to smooth out instances like 'notgood' or 'not bad'. When negation as an explicit feature didn't help, we thought of experimenting with bigrams. However, they happened to be too sparse in the data and the overall accuracy dropped in the case of both NB and MaxEnt. Even collapsing the individual words to equivalence classes did not help. Bigrams however happened to be a very sparse feature which can be seen in the outputs with a lot of probabilities reported as 0.5:0.5.For context: @stellargirl I looooooooovvvvvveee my Kindle2. Not that the DX is cool, but the 2 is fantastic in its own right. Humans [0.5000] Cyborg [0.5000] 3. Negate as a features Using the Stanford Classifier and the base SVM classifiers we observed that identifying NEGclass seemed to be tougher than the POS class, merely by looking at the precision, recall and F1 measures for these classes. This is why we decided to add NEGATE as a specific feature which is added when “not” or „n‟t” are observed in the dataset. However we only observed a increase in overall accuracy in the order of 2% in the Stanford Classifier and when used in conjunction with some of the other features, it brought the overall accuracy down and so we removed it. Overlapping features could get the NB accuracy down, so we were not

very concerned about the drop with NB. However it didn't provide any drastic change with OpenNLP either. 4. Part of Speech (POS) features We felt like POS tags would be a useful feature since how you made use of a particular word. For example, „over‟ as a verb has a cyborg connotation whereas „over‟ as the noun, would refer to the cricket over which by itself doesn‟t carry any cyborg or humans connotation. On the Stanford Classifier it did bring our accuracy up by almost 6%. The training required a few hours however and we observed that it only got the accuracy down in case of NB Handling the Bots Class In the previous sections, bots was disregarded. The training and test data only had text with humans and cyborg s. In this section, we explore what happens when bots is introduced.

Retweet Rule

Twitter's Retweet feature (aka RT) is one of the most influential features of Twitter for getting the word-of-mouth circulated quickly to so many users. RT is a very powerful tool such that when a user searches for a tweet or receives one, they can choose to share it with their followers via this feature. Once a tweet is retweeted it shows up to all the original sender's followers. After identifying a relevant tweet, this rule further decides whether it should be shared in the form of RT or not.

RT Rule Implementation.

```
 (defrule action-retweet
(twitter-user
(language "en")
(screen-name ?screenName))
(raw-tweet-info
(id ?tweetID)
(text ?tweetText))
(test (and (contains-retweet-keywords
?tweetText)
(contains-article ?tweetText)))
=> (assert (recruitment-action
(action "retweet")
(tweetID ?tweetID) )))
```

Types of Hashtags Found

The following are the various types of hashtags that were generated from the small dataset of 25 tweets. For the sake of enumerating some knowledge types hashtags can produce, we set the minimum support and minimum confidence to 2% and 100% respectively. This low support guarantees the inclusion of many hashtags regardless of the how frequent they are in the given rules. The network mining described in Section 4.4 explains how we prune insignificant rules. Figure 4.3 shows the association rules generated from a 25 tweet dataset by analyzing the the keywords when hangtags are incorporated. In the following section, we present hashtag types, meanings, and significance in the Twittersphere: [noitemsep,nolistsep]Activity– Five different discovered hashtags suggested an activity type: The first three (#TTOT which is an acronym for Travel Talk on Twitter, #RTW an acronym for Round the World, and #travel) which were linked to the MeSH Smoking concept and travel.

In information retrieval with binary classification, precision (also called positive predictive value) is the fraction of retrieved instances that are relevant, while recall (also called sensitivity) is the fraction of the relevant instances that are retrieved. Precision and recall are therefore based on understanding and measuring relevance.

In simple terms, high accuracy means that an algorithm returns significantly more relevant than irrelevant results, while a high recall means that an algorithm has yielded the most relevant results.

The most important category measurements for binary categories are:

Precision	Recall	F Measure
$P = TP/(TP + FP)$	$R = TP/(TP + FN)$	$tp + tn/tp + tn + fp + fn$

Classification	Precision	Recall	FScore
Naive Bayes	99.78	98.23	98.10
C45	95.78	95.23	94.90

Table 4.4 Classification Results

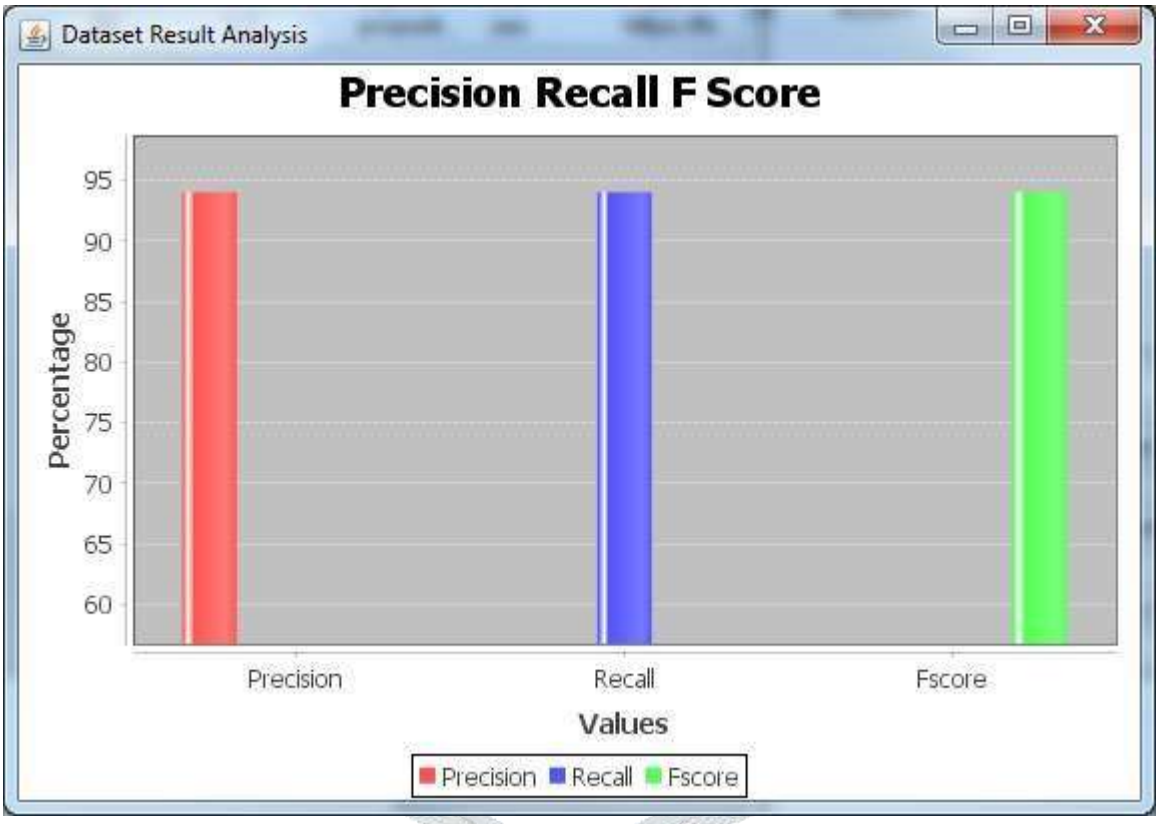


Figure 4.1: Result Classification for C45 Classification

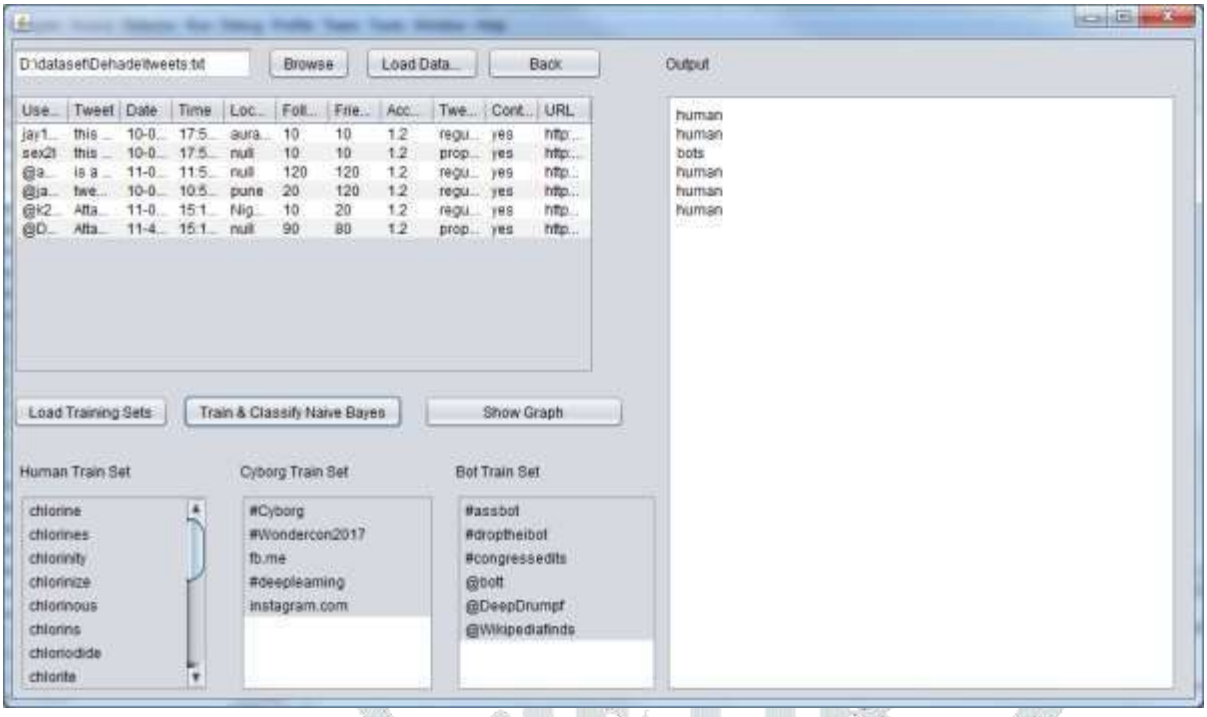


Figure 4.2 Naive Bayes Classifier

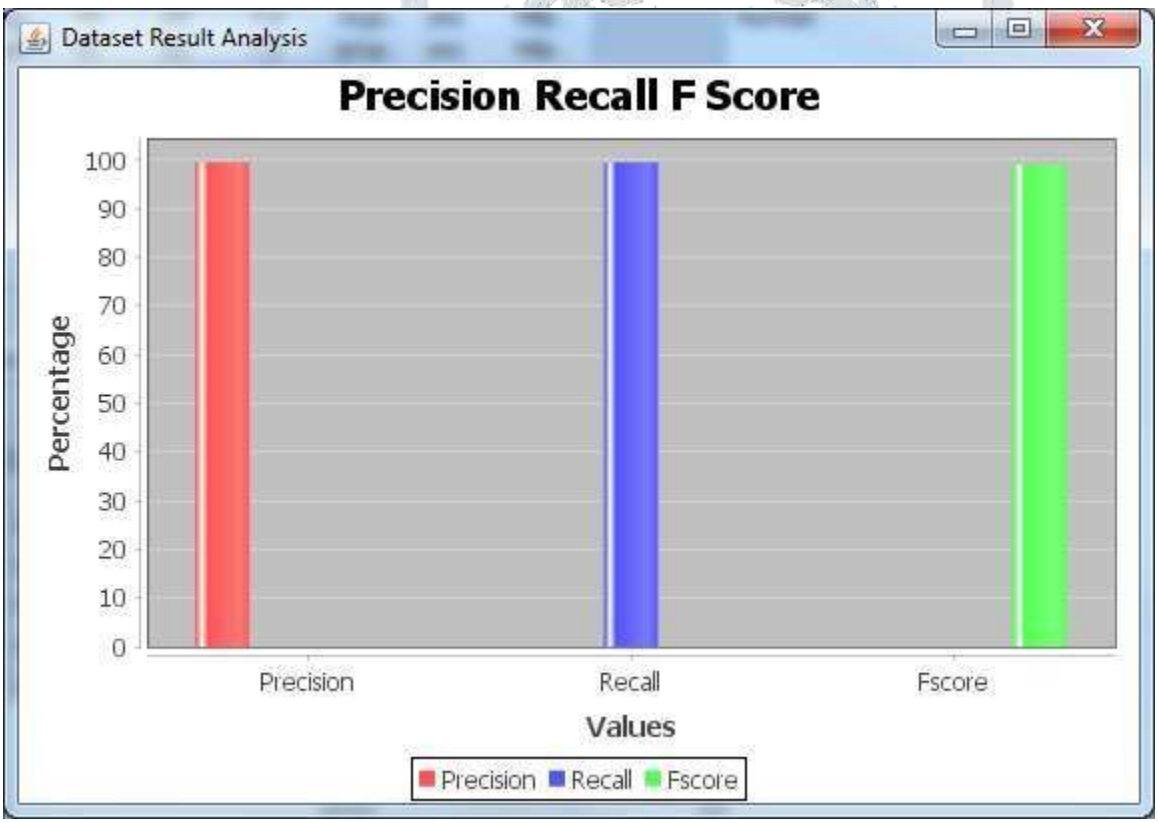


Figure 4.3 Results for Naive Bayes Classification

Conclusion

In order to obtain the most relevant parameters, a large number of quality sentences will be analyzed to determine the characteristics of teenagers and adults, based on writing style and user profiles and profiles.

Some parameters have been removed because they do not affect the final classification results, making it clear that they should not consider or use the Word2Vec format as a learning algorithm for machines that provide the best results for age group classification. The importance of considering the profile of users, data in measuring the intensity of that feeling, has been claimed in many studies.

Social networks do not provide user information or users limited personal information. In these cases, the proposed model for age group prediction is very important in improving the efficiency of emotional intensity measurements. The models we offer have better situations that are not available. In addition, the proposed model can work with other confidence measurements.

References

[1] Guimaraes, Rita & Rosa, Renata & De Gaetano, Denise & Rodriguez, Demostenes Zegarra & Bressan, Graca. (2017). Age Groups Classification in Social Network Using Deep Learning. IEEE Access. PP. 1-1. 10.1109/ACCESS.2017.2706674.

[2] R. G. Guimaraes, D. Z. Rodr ˜ ´ıguez, R. L. Rosa, and G. Bressan, "Recommendation system using sentiment analysis considering the polarity of the adverb," in 2016 IEEE International Symposium on Consumer Electronics (ISCE), Sao Paulo, Brazil, Sep 2016, pp. 71–72.

[3] C. Peersman, W. Daelemans, and L. Van Vaerenbergh, "Predicting age and gender in online social networks," in Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents. Glasgow, Scotland, UK: ACM, Oct 2011, pp. 37–44. [Online]. Available: http://doi.acm.org/10.1145/2065023.2065035
[4] J. Van de Loo, G. De Pauw, and W. Daelemans, "Text-based age and gender prediction for online safety monitoring," Computational Linguistics in the Netherlands, vol. 5, no. 1, pp. 46–60, Dec 2016.

[5] R. M. Filho, J. M. Almeida, and G. L. Pappa, "Twitter population sample bias and its impact on predictive outcomes: A case study on elections," in IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Paris, France: ACM, Aug 2015, pp. 1254–1261.

[6] D.-P. Nguyen, R. Gravel, R. Trieschnigg, and T. Meder, "How old do you think i am? a study of language and age in twitter," in Seventh International AAAI Conference on Weblogs and Social Media. Palo Alto, CA, USA: AAAI Press, Jul 2013, pp. 439–448.

[7] L. Sloan, J. Morgan, P. Burnap, and M. Williams, "Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data," PloS one, vol. 10, no. 3, pp. 1–20, Mar 2015.

[8] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker, "Effects of age and gender on blogging," in AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, Stanford, CA, Mar 2006, pp. 199– 205.

[9] S. Goswami, S. Sarkar, and M. Rustagi, "Stylometric analysis of bloggers age and gender," in International AAAI Conference on Web and Social Media, San Jose, California, May 2009, pp. 214–217.

[10] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler, "Mining the blogosphere: Age, gender and the varieties of self-expression," First Monday, vol. 12, no. 9, pp. 214–217, May 2007.

[11] S. De Jonge and N. Kemp, "Text-message abbreviations and language skills in high school and university students," Journal of Research in Reading, vol. 35, no. 1, pp. 49–68, Oct 2010.

[12] D. A. Huffaker and S. L. Calvert, "Gender, identity, and language use in teenage blogs," Journal of Computer-Mediated Communication, vol. 10, no. 2, pp. 01–24, Jun 2005.

[13] L. A. S. Shapiro and G. Margolin, "Growing up wired: Social networking sites and adolescent psychosocial development," Clinical child and family psychology review, vol. 17, no. 1, pp. 1–18, Mar 2014.

[14] S. Rosenthal and K. McKeown, "Age prediction in blogs: A study of style, content, and online behavior in pre-and post-social media generations," in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. Portland, Oregon: Association for Computational Linguistics, Jun 2011, pp. 763—-772.

[15] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in twitter," in Proceedings of the International Workshop on Search and Mining User-generated Contents. Toronto, Canada: ACM, Oct 2010, pp. 37–44.

[16] F. Barbieri, "Patterns of age-based linguistic variation in american english," Journal of sociolinguistics, vol. 12, no. 1, pp. 58–88, Jan 2008.

[17] L. Zheng, K. Yang, Y. Yu, and P. Jin, "Predicting age range of users over microblog dataset," International Journal of Database Theory and Application, vol. 6, no. 6, pp. 85–94, Oct 2013.

[18] R. L. Rosa, D. Z. Rodr´ıguez, and G. Bressan, "Music recommendation system based on user's sentiments extracted from social networks," IEEE Transactions on Consumer Electronics, vol. 61, no. 3, pp. 359–367, Oct 2015.

[19] R. L. Rosa, D. Z. Rodriguez, and G. Bressan, "Sentimeter-br: Facebook and twitter analysis tool to discover consumers sentiment," in The Ninth Advanced International Conference on Telecommunications, Rome, Italy, Jan 2013, pp. 61–66.

[20] R. L. Rosa, D. Z. Rodr´ıguez, and G. Bressan, "Sentimeter-br: A social web analysis tool to discover consumers' sentiment," in IEEE 14th International Conference on Mobile Data Management, Milan, Italy, Jun 2013, pp. 122–124.