Classifying Twitter Data using Deep Learning Technique

[1]Shantanu Danekar, [2]Prof A.H. Rokade

[1]ME Student, [2]Professor
Computer Engineering Department,
Shreyash College of Engineering and Management, Aurangabad

***Abstract*: Online social networks have a lot of information. But often people do not provide personal information, such as age, gender and other demographic data, although the confidence analysis uses such information to develop useful applications in people's daily lives. But there is still a failure in this type of analysis, whether by the limited number of words contained in the word dictionary or because they do not consider the most diverse parameters can influence feelings in sentences; therefore, more reliable results will be obtained if considering user profile data and user writing style. This research shows that one of the most relevant parameters contained in the user profile is the age group, which shows that there is normal behavior among users of the same age group, especially when these users write about. With the same topic detailed analysis with 7000 sentences has been conducted to determine which features are relevant, such as the use of punctuation, number of characters, sharing of media, other topics, and which ones can ignore the age group classification. Different learning machine algorithms have been tested for the classification of adolescent and adult groups and the Deep Convolutional Neural Network (DCNN) has the best performance with accuracy up to 0.95 in the validation test. Must In addition, in order to verify the usefulness of the proposed model for age group classification, it is implemented in the Sentiment Metric (eSM) that has been improved. In performance audits, subjective tests are performed and eSM with the proposed model arrives. Mean Square root error and Pearson's correlation coefficient of 0.25 and 0.94, respectively, are more efficient than the eSM indicators when no age group information is specified.**

***Keywords:* Tweets, NLP, Twitter, Deep Learning**

Introduction

Users spend time browsing websites, e-commerce, reading news about sports, journalism and entertainment, including expressing their opinions and feelings in the form of comments on social networks about a variety of topics. These comments can be analyzed to assess customer satisfaction as a very useful information for service providers and product suppliers. Goldsmith and 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 between each other. There are many applications that are used that use confidence analysis, such as psychological disease detection [2], false profile detection [3] to prevent criminals from attracting new victims. [4] Predict 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 which can be collected and analyzed. However, informal and short sentences with a large number of languages [3] require some parameters to be improved to analyze data. Among them is the age that can directly influence the final sentences of sentences [6], [7]. The general characteristics found in each period of life are considered in this type of analysis. Especially those characteristics will be clearly different in teenagers and adults. It is important to note that in some social networks, the age of the user cannot be used either on the social network itself or by the user for reasons according to the discretion. Therefore, determining the method for predicting the age of users is therefore relevant in analyzing confidence. 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 accuracy up to 80.32% [9], [10]

De Jonge et al. [11] Check the abbreviations of text messages in high school and university students, such as emoticons, slang, lengths, text and spelling mistakes. Huffaker and faculty [12] examine the use of teenage language in blogs. They concluded that the most commonly used language is abbreviations and emoticons, Shapiro and the faculty. [13] Study the frequency at which teens write on social networks. 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 slang sentence length, the use of the Uniform Resource Locator (URL)) to share media information. , The number of people who follow, number of followers, the total number of tweets posted on social networks and topics that are relevant to increase courage Expression and accuracy in classifying age groups Some of these parameters are used in other applications such as slang, emotion icons, and sentence lengths [14] - [16]; But they do not consider parameters such as punctuation, URLs, people that users follow, 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. In addition, our research also considers which parameters can be discarded while classifying age groups, such as references used. @ Symbols and hashtag and message sharing Education levels are not considered in this work because they have been tested and have low accuracy. It is worth noting 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 education has been conducted using social networks Twitter, but can be extended to other social networks because the parameters are often the same. In this research, to classify adolescents and young adults,

different machine learning algorithms have been tested and neural networks. Convolutional (DCNN) to the best performance In addition, in order 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 conducted and results Compare with the following confidence indicators SentimeterBR2 [19], [20], eSM, without considering the format proposed eSM with the proposed model and eSM that considers the actual age group. These results show the relevance to determine age group parameters and benefits. Of the proposed model in confidence analysis.

Literature Review

In this section, the main study of the influence of certain parameters in the classification of age groups has been observed. There is also a discussion of confidence analysis and machine learning algorithms. In the sentiment analysis, some studies have been cited to emphasize that the user's age data is an important parameter in improving the efficiency of measuring the intensity of feelings.
A. The relationship between age groups and the nature of writing.
Psychology shows the difference in the behavior of people in different ages. [21], [22] In general, teens don't care about their privacy. [23] and they post and publish a number of information. A lot on social networks online It can be considered a teenager as a person of up to eighteen. That is when they reach the age of majority. But for cultural reasons, some countries use between 13 and 19 years. [21] Age information is not available in some social networks, such as Twitter. After checking that this information can actually change the results of many analyzes, research Something [3], [17], [24] trying to predict it.
One strategy used is to search for descriptions in profiles with expressions "X years", "I have X years" or "I use X years" where X represents the age of the user. However, it has been verified that on Twitter, the age profile in profile details is not a common habit. [25] Therefore, these studies will not give reliable results. It is common among young users of social networks to discuss other topics that occur in their daily lives, affecting their real world. [10] Topics such as school relationships And friends often live in this age group. [12].
Mature users relate to their own images. Then they will be more careful with the comments they write and those who can read [26]; Therefore, it is possible to find sentences that have more positive feelings, not using self-referencing, to use less denial. [8] Therefore, using slang becomes less [27]. Less about yourself can be proved by the time users online. In adulthood, users have more commitments throughout the day and teens spend more time with online media than hours per day. Then, for teenagers, social networks became an important tool to express their opinions on the world. [13]
In addition to identities of traditional identity in adults, such as topics of religion, ideology, politics and work; Adults also use online media to express opinions. Mature users are familiar with attaching images, videos or sharing links of other pages that will fill the information that begins in tweets. [28] Between these two groups, Twitter users under the age of thirteen are not considered in this research because many online social networks require users to Can be used for at least thirteen years to be able to apply Therefore, a group of teenagers consisting of users aged 13 to 20 years, also known as teenagers and adults, consists of all users aged 20 and over.

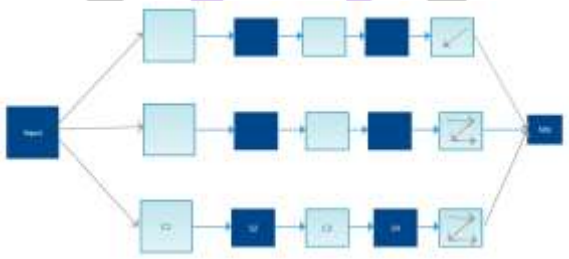
In this context, it is common to find studies that analyze the opinions of e-commerce websites and social networks by separating the feelings of that comment. There are many studies on confidence analysis. But most do not consider user profiles such as metrics Sentimeter-Br2 That is based on the dictionary, which each word has a positive or negative value of confidence This indicator considers n-grams, adverbs and no words to stop. Differences, values, confidence depend on the verbal words in which the past verbs have less confidence than the verbs of the present time. Sentimeter-Br2 That is located on Sentimeter-Br [29].
The study described in the following describes the impact of user profile data on confidence analysis with the aim of increasing the effectiveness of the confidence indicators. 1) ANEW is a study that considers data related. Users for confidence analysis [30] have studied whether the existence or absence of gender data will affect the end result. 2) SentiWordNet is a measure of connectivity. Users are used to classify polar confidence automatically aims to assist in the analysis of social media. [31].
The author analyzes how important it is to consider the user profile and later adjusting the focus of the study to cover various language analysis. Users, including gender, education level, geographic location, and age, eSM is a relationship of confidence indicators that use lexicon dictionary, Sentimeter-Br2. With the editing factor based on the user's profile information, eSM format of Fi sentences set by (1). Please note that there is an assumption that that age information exists in the user profile.
In addition to these work, others [33], [34], report that in analyzing rigorous sentiment, it is important to have indicators for the examination of the irony because this can reverse the Convinced sentences In both works, the age of the user can influence how to rank the sentences ironically. However, on Twitter, age information is confidential and not mandatory.
Similarly, [12] the author found that teenagers behave differently in the online environment and it is possible to observe certain characteristics in the writing style, such as topics that affect their world. Setting up posts about topics that do not refer to themselves and dealing with more positive information may be the characteristics of mature users [26]; While using slang is often found in sentences posted by teenage users. [27] Moreover, the need to attach the media that represents the content mentioned is the characteristics of adult users. [28].
Therefore, this work is intended to group the characteristics mentioned earlier [12], [27], [28] and others proposed in this research, such as the use of punctuation, abbreviations, symbols that express emotions and characteristics of Writing style In addition to other user information such as history, tweets, number of followers and the number of people he or she follows.
There are many machine learning models [35] that cover basic methods such as linear regression and tree models, including more sophisticated methods such as artificial neural networks or vector support. In general, machine learning is not limited to only one data analysis. But compare many models and choose the model that has the best predictive accuracy Machine learning area, also known as pattern recognition or data mining [36], involves the separation of patterns in large data sets. Often the goal is to predict the correct response variable, for example, age groups based on one or more prefixes, such as writing style.

Qualitative methods are often not considered in research about confidence analysis. [37] There is a need to filter or delete meaningless sentences that are considered the most vocal in the data warehouse by Twitter.] - [40] which has a lot of information Many times in the early stages of data analysis, it is necessary to specify the main characteristics or patterns of the sample and this work is done by the experts themselves. In this context, research is conducted on a large amount of data and is filtered and covered in a population, regardless of specific people. Then even more personal information such as: "I miss you home" or "I'll start reducing carbohydrates." Do not share the user's personal information in the results. In the end, the goal is to draw patterns that exist in the manner of expression from each age group and not just in some cases.

In order to be able to work with a large amount of data and achieve the desired classification, the machine's learning algorithm is used which can provide high accuracy results. [41] For the algorithm instance that Using decision trees (J48), vector support (SMO) or artificial neural networks, the use of deep Leaning increases in many ways, which has occurred in recent years, such as images [42], [43]. And Strasbourg voice The deep leaning algorithm allows the computational model that consists of multiple processing layers to learn the representation of data with multiple levels of abstraction. Recently, deep learning methods have been applied to text classification. [44], [45] and the algorithm has excellent results for the classification of text patterns [46] - [49]

Deep learning is often interpreted in terms of the universal approximation theorem [50] or the probability inference [51]; The approximation theorem defines a class of universal approximations, which refers to the ability of the neural network of feeding directly with a single, limited-size mystical layer for approximately continuous functions. Interpretation of probability comes from machine learning, including inference, as well as optimization concepts such as training and testing related to adaptation and general characteristics.

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 Deep Convolution Neural Network (DCNN).

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

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

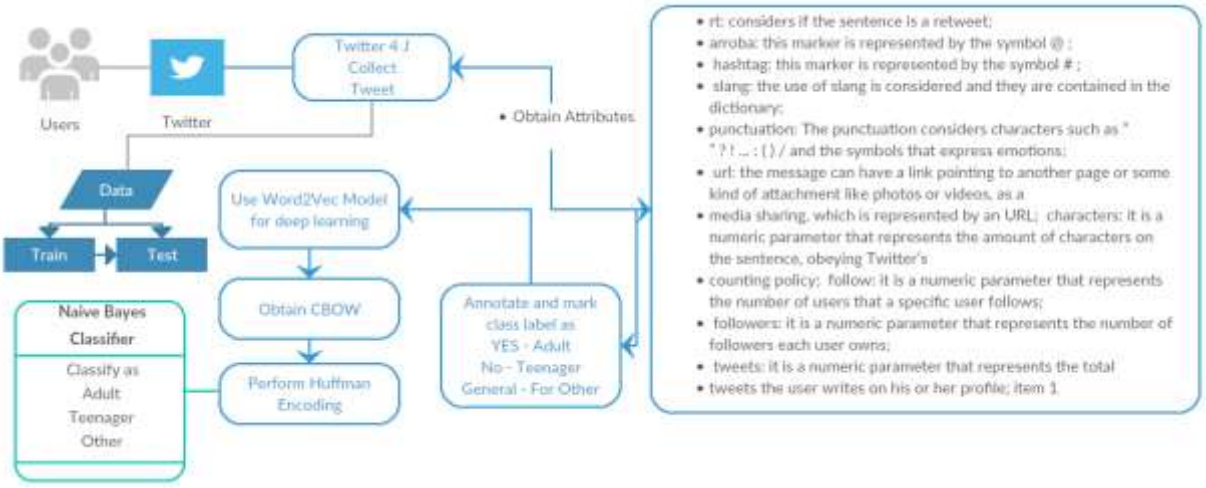
Where:

- 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

Twitter4J:

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.

Natural Language Processing (NLP):
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.

Feature Extractors
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

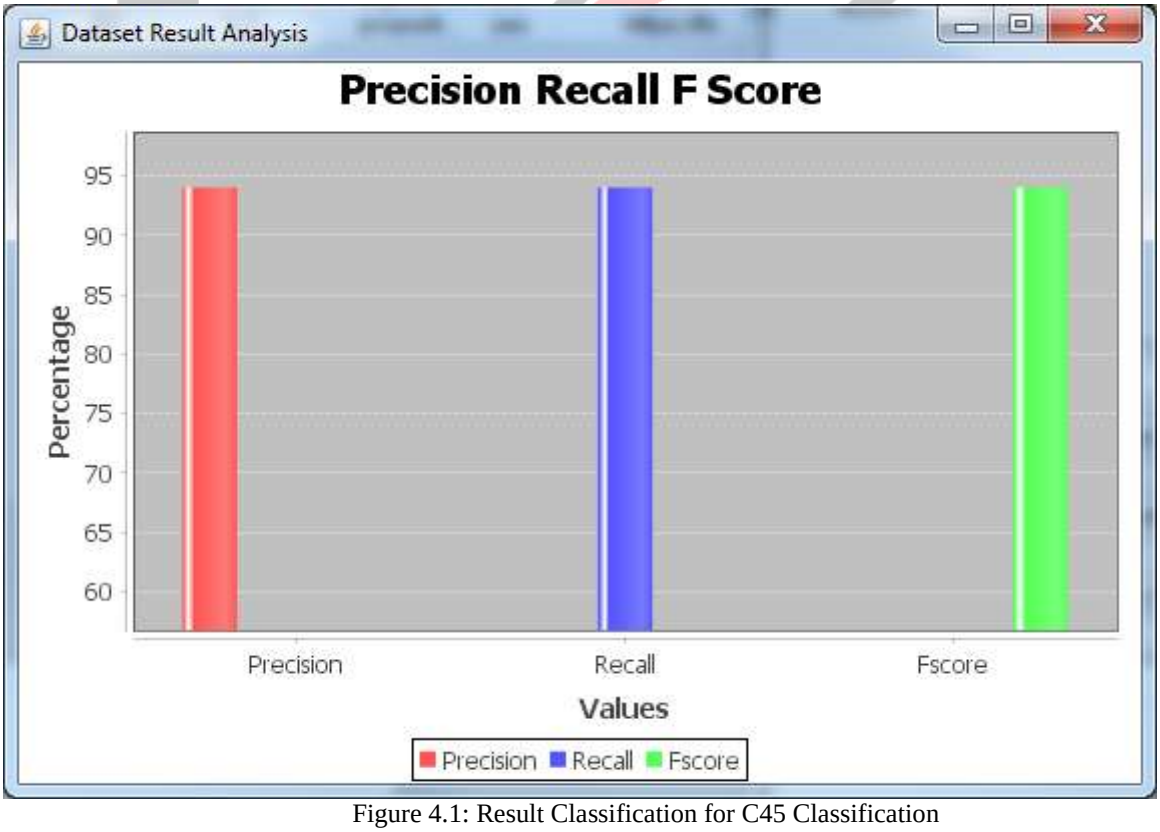


Figure 4.1: Result Classification for C45 Classification

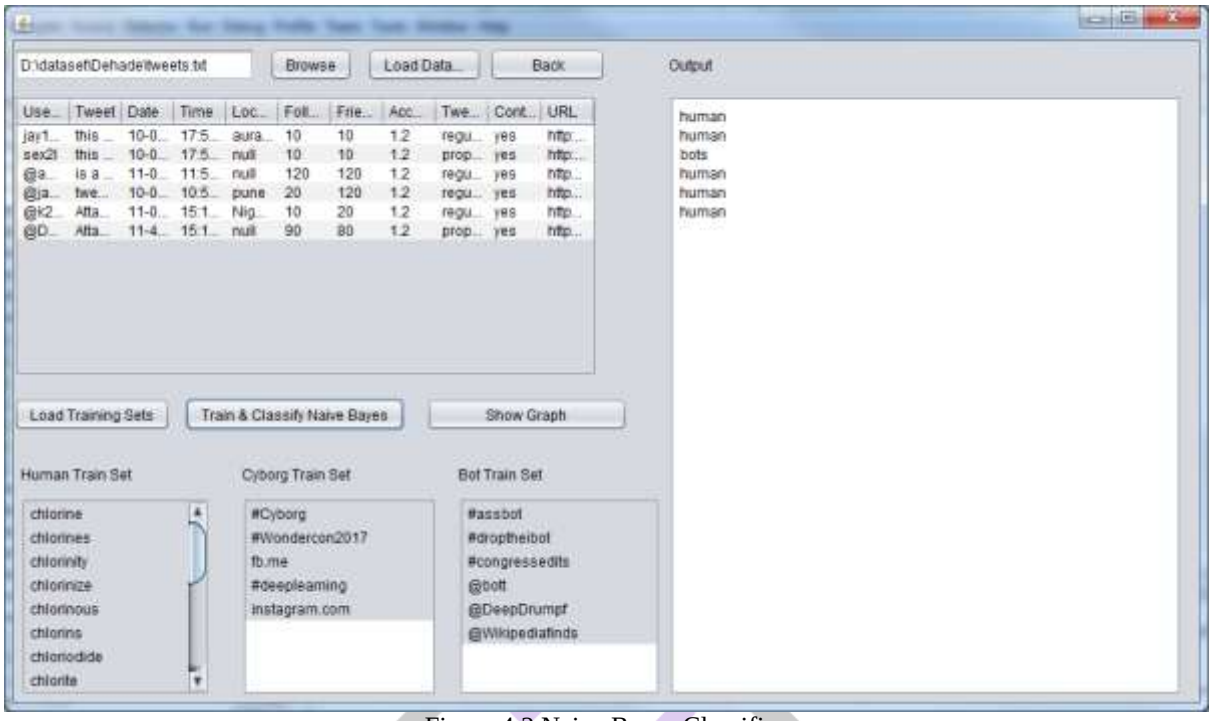


Figure 4.2 Naive Bayes Classifier

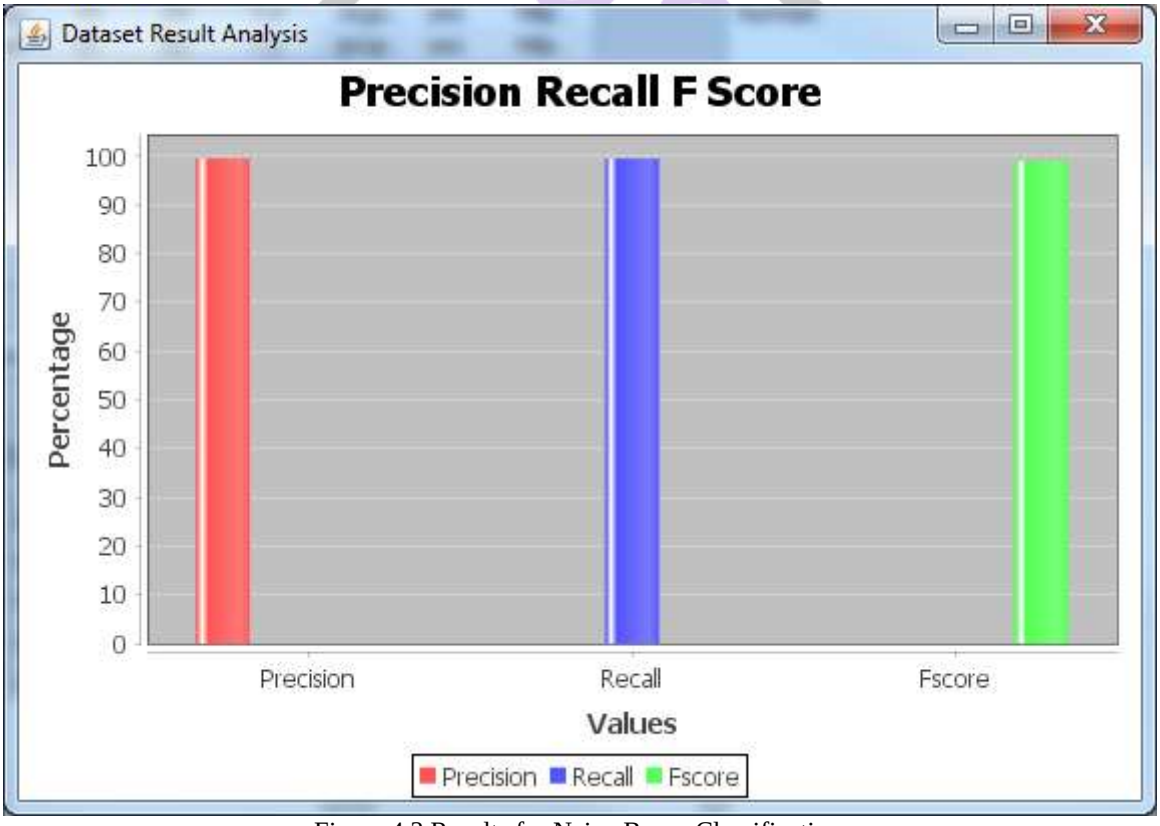


Figure 4.3 Results for Naive Bayes Classification

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.