EAST WEST UNIVERSITY

An Incremental Approach of Classifying Age Group Using Text Analysis on Blog Data

Presented by

Md. Manir Haider Helalee

Id: 2016-1-50-006

Obayedur Rahman

Id: 2016-1-50-019

Md. Tanver Hasan

Id: 2016-1-50-007

Supervised by

Dr. Mohammad Arifuzzaman

Assistant Professor

Department of Electronics and Communications Engineering

East West University, Dhaka

This Project submitted in partial fulfilment of the Requirements for the Degree of Bachelors of Science in Information and Communications Engineering

To the

Department of Electronics and Communications Engineering
East West University, Dhaka, Bangladesh

Declaration

We, hereby, declare that the work presented in this thesis is the outcome of the investigation performed by us under the supervisor of Dr. Mohammad Arifuzzaman, Assistant Professor, Department of Electronics and Communications Engineering, East West University. We also declare that no part of this thesis/project has been or is being submitted elsewhere for the award of any degree or diploma.

Countersigned

.....

(Dr. Mohammad Arifuzzaman)

Supervisor

Assistant Professor,
Dept. of ECE,
East West University

Signature

.....

(Md. Manir Haider Helalee)

2016-1-50-006

Signature

.....

(Obayedur Rahman)

2016-1-50-019

Signature

.....

(Md. Tanver Hasan)

2016-1-50-007

Letter of Acceptance

This project entitled “An Incremental Approaches of Classifying Age Group Using Text Analysis on Blog Data” submitted by Md. Manir Haider Helalee, ID: 2016-1-50-006, Obayedur Rahman, ID: 2016-1-50-019, Md. Tanver Hasan, ID: 2016-1-50-007 to the Electronics and Communications Engineering Department, East West University, Dhaka-1212, Bangladesh is accepted as satisfactory for partial fulfillment of requirements for the Award of Degree of Bachelors of Science (B. Sc.) in Electronics and Communications Engineering on December, 2019.

Chairperson

. . . . . .

(Dr. Mohammed Moseeur Rahman)

Chairperson and Assistant Professor,

Department of Electronics and Communications Engineering,

East West University

Supervisor

. . . . . .

(Dr. Mohammad Arifuzzaman)

Assistant Professor,

Department of Electronics and Communications Engineering,

East West University

Abstract

Since the dawn of civilization, people are using the writing method to express their thoughts and views. To expose one's feelings it is the best way till now. Social network and many different blogs have a large amount of data, but people don't provide their personal data such as age and other demographics. Age groups classification from text analysis has become a leading context for scientific and commercial market research in the field of machine learning. Currently, it's a more prominent research field of English language processing system as there is few researches works regarding text analysis for this language. There are still failures to identify perfect age group because they do not consider the most important parameter which can influence the overall result. The main objective of this research is to develop systems that which word are more frequent in a particular age group. Different machine learning algorithm is used for the classification of the teenager and adult group. Almost 100k sentence was performed to determine which parameter is relevant. Logistic Regression with TF-IDF had the best performance reaching a precision .80 in the validation test. To make the mechanism more efficient and accurate unigram method has been implemented. Several techniques have been integrated for data collection and data processing to make the system more reliable and flexible. Adequate instances and experiments are also provided to describe the methodology for both approaches.

Acknowledgment

As it is true for everyone, we have also arrived at this point of achieving a goal in our life through various interactions with and help from other people. However, written words are often elusive and harbor diverse interpretations even in one's mother language. Therefore, we would not like to make efforts to find the best words to express my thankfulness other than simply listing those people who have contributed to this thesis itself in an essential way. This work was carried out in the Department of Electronics and Communications Engineering at East West University, Bangladesh.

First of all, we would like to express my deepest gratitude to the Almighty for His blessings on us. Next, our special thanks go to our supervisor, Dr. Mohammad Arifuzzaman, who gave us this opportunity, initiated us into the field of Machine Learning, and without whom this work would not have been possible. His encouragements, visionaries and thoughtful comments and suggestions, unforgettable support at every stage of our B.Sc. study were simply appreciating and essential. His ability to muddle us enough to finally answer our own question correctly is something valuable what we have learned and we would try to emulate if ever we get the opportunity. I also want to thank all faculty members and staffs of Department of Electronic and Communications of East West University, who have shown me their constant support and friendship in various ways, directly or indirectly related to our academic life. We will remember them in our heart and hope to find a more appropriate place to acknowledge them in the future.

Md. Manir Haider Helalee

Obayedur Rahman

Md. Tanver Hasan

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Now - a – days, by increasing the use of internet, peoples are spending their lot of time in the social site such as Facebook, Twitter, Instagram, Myspace, Hive's, Bebo and Net log, Snapchat, LinkedIn, YouTube, WhatsApp etc. Text Analysis has become a topic of much interest and development for research area as it has many practical and empirical applications. Since openly and secretly available data over the web is continually flourishing, an enormous number of writings uncovering conclusions are accessible in various web journals and different articles in online diary, web-based life, and item audit locales. Users are expressed their opinions and sentiment through reviews, comments and updates by using those sites. Peoples are sharing their feelings, emotions by writing message or share post in social site. Now, various companies are using this site to know the customer satisfaction with the help of this post, comments and messages. They share their products on that site and try to find out the customer of a particular age so that they can provide better product depend on the customer need. There is a common characteristic that it is easy to give false identity such as false name, age, gender, and people's location where they are living. Through the facilitation of text analysis system, this unstructured and scattered information could be automatically transformed into structured data, and by sensing this actual data, opinions can be extracted. Besides this, with the current progress of machine learning and text mining techniques, the competency of algorithms to analyze the text has progressed numerously. Because of that, the conduct of text analysis is developing day by day in the field of product analysis, social media monitoring, brand monitoring, workforce analysis, business analysis, market research and analysis and so on. Consequently, text analysis has turned out an important research area over the last decade. In this research we classify the age group by classifying word. We classify the teenager and adult age group using different learning algorithms.

1.2 Overview and Motivation

We choose text analysis to easily find the age from the microblog data. Here, we used different types of algorithm such as random forest, linear SVC, TF-IDF, Linear regression, Naïve based algorithm and many one to find the accurate age of the user. We also used cleaning for clean the unexpected data (special character, tag etc.) so that we can find the better accuracy.

1.3 Why Text Analysis?

Text analysis is the process that describes unstructured text and eliminates related information, and then it will transfer to work for business sector, Social Media Monitoring, Market Research and Analysis. Now - a- day's text analysis is broadly used for finding internal site from various types of social media reply, post, comments, survey responses, and product reviews, and creating data-driven idea or method. In the world, we propagate 2.5 quintillion bytes of information each and everyday, text analysis most important and one of the most successive tool for creating that type of data.

1.4 The Reason behind Choosing Text classification

Text analysis is a very important supply that is applied to the voice of client materials. Text analysis will be extracted, evaluated and identified from the reviews and survey responses, online and social media, and health-care materials. Basically, it deals with the user grasp regarding individual reality. Peoples are sharing their concepts, thinking, and opinion with a brief message via different diary and social networking platform that brings new scope for developing artistic client service solutions. Text explains analyses of information that is extracted by different technologies and data processing techniques. Now-a-days most of the peoples everywhere the planet are connected with social network platform like Facebook, twitter, LinkedIn, reddit wherever users will categorical their opinion on specific product, service, brand, political, sports etc. Insights from the text analyses the corporate have gotten plan concerning the product and analyze the unstructured text, then structured this data by using different types of algorithm. Political parties are often famed what quantity folks support their work, Social organizations and NGOs firms will grasp people's opinion by questioning. These text analyses are evaluated by classifying the polarity of a text at a document.

1.5 Application and benefit of The Text Analysis

Business Analysis: Text Analysis can be utilized in brand observing, item investigation, client assistance which has a significant impact in business examination. Trough Text examination individuals' conclusions can be identified effectively in regards to a specific item or brand. Indeed, even it tends to be utilized to check out the potential strings developing internet with respect to one's matter of fact. Thusly, by utilizing this data, any association can be proactive and can change or improve their system in industry all the more rapidly.

Social Media Monitoring: These days, loads of individuals are utilizing diverse online life. In this way, web-based life has become a decent wellspring of data. Consequently, to reveal the concealed data from the writings in web-based life has gotten vital for such a large number of divisions. With the assistance of trend setting innovation and by utilizing content examination ideas, internet-based life can be checked rapidly and can likewise increase important data over yonder. State for instance, for a specific post on Facebook or Twitter by breaking down the remarks of the crowds the content of group of spectators for that particular post can be effectively distinguished.

Market Research and Analysis: The idea of text analysis is currently applying in statistical surveying examination. To know the new pattern, individuals' response over any new items, to find out about new business approach or to make a correlation with various brand items dependent on client's audit content has become a basic factor here.

In Corporate Network: Text analysis can likewise be applied in the corporate system. For instance, by utilizing this idea to the email server, messages can be effectively observed. As of late this feeling examination framework has used to recognize various types of messages like significant mail, limited time or spam.

1.6 Scope of Text Analysis

Text analysis can be applied at the different level of scope. They are –

• **Document/ Article Level:** At the point when Text analysis is applied to extricate the feeling from a total record of content or article that is called report/article level of extension.

• **Sentence Level:** At the point when a text analysis framework works for assuming from a only one (single) sentence.

• **Sub Sentence Level:** At the point when Text analysis separates and guess from a short part or an expression of a sentence.

In this research paper sentence and document level of magnitude for determining emotions has been implemented for both proposed approaches.

1.7 Methods of Text Analysis

There are various type of algorithms and methods of classification techniques to implement text analysis. Those techniques can be categorized as- rule-based system, automatic system, and hybrid mechanism. For Text most of the research work on sentiment analysis has been done by using Rule-Based systems that used a variety of inputs (NLP techniques, Use other resources like Lexicons). However, this examination work extricates feeling by utilizing Automatic frameworks which are inverse of the standard based frameworks and don't trust in the physically created principles. This framework depends on AI methods to gain from information.

1.8 Conclusion

With the help of advanced artificial intelligence, machine learning, and text mining techniques the field of text analysis is growing day by day. In future, this concept of sentiment analysis could be applied in so many practical fields. For these lots of research work and new methods are needed to invent to get more accurate results in this sector. For English and other languages there exists a lot of research works regarding this topic. But, for Bangla language, more research work and their practical applications are needed in this field. Therefore, because of its so many rich applications, text analysis has become a popular topic in research and business analysis field.

1.9 Organization of Thesis

The rest of the part of this paper is organized by the following sections. Chapter 2 provides gradually a short description of some previous works related to our topic of interest. Chapter 3 gives a simple view of the overall working procedure. In Chapter 4, provides information regarding dataset (learning materials) preparation. In Chapter 5, we briefly describe the methodology for proposed methods with proper instances. In Chapter 6, gives a descriptive view of the result and analysis of our works. In Chapter 7, we conclude the paper and provide some directions for future work

Chapter 2

Related Work

2.1 Introduction

Nowadays, Text Analysis has become a hot topic for researchers because of its practical applications and lucrative benefits in so many fields. With the help of many advanced classification algorithms in machine learning and text mining techniques, this text analysis process is developing day by day. A lot of research work has been done on text analysis by using so many advanced techniques. There exist a few researches works regarding text analysis for the age predicting. Therefore, Text analysis has become a popular field for researchers as this field is still at a growing stage for the age predicting.

2.2 Motivation of this Research Work

To the extent the researcher knows none of the examination chip away at text Analysis from age predicting information has been done to decide the age for utilizing the various kinds of classifier. All the research works takes a shot at content investigation for predicting the age have been finished by utilizing AI and NLP systems. As there are best look into works for predicting the age on text analysis by utilizing propelled AI procedures, so it has become a huge requirement for predicting the age.

2.3 Existing Research Work Regarding Text Analysis for Age Group Classification

Some previous works on Text analysis are narrated here. In paper [1], the author proposed that one of the most important parameters contained in the client profile is the age gathering, demonstrating that

there are commonplace practices among clients of a similar age gathering, specifically, when these clients expound on a similar point. They are the different type of age and character. There is a research on the post that user share their feelings and expression in their social site [2], [3]. In paper [4], researcher said that teenagers, they can't know the privacy of social site and they post any types of opinion in social site. From those post people know their situation and their valuable information. In paper[2], author research on age that some country call that children are teenagers if their age is 18 and some countries are called children's are teenagers if the range of the age are used the range between 13 and 19 years. In paper [5], the author utilized the idea of term frequency and inverse document frequency (TF and IDF) value to show signs of improvement arrangement, and they accomplish a progressively precise outcome by removing the various highlights of positive, negative or nonpartisan expressions of text examination. In paper [6], the creator proposed a feeling following framework on theme or occasion was completed by utilizing sense-based influence scoring methods by utilizing SentiWordNeT for Text analysis. There is an another research regarding on the most commonly teenagers are post their feelings and their opinion on Facebook, tweeter, Instagram and the researcher are analyze this information and then use this text by using different types of methods using AI and machine learning [7],[8],[9]have worked on trying to predict the age. In the paper [10], the author wrote on the topics such as relationships, school and friends are more convenient in this age group. Another study discuss about the adult user are know the privacy of the social site and they also concern their post and they can't post that type of post so that they can fall in serious danger [11]. In paper [18], the author proposed that they can found positivity and negativity of a topic (Samsung, IPhone) by using text analysis. In paper [19], author proposed that we can locate the great outcome by utilizing content examination when we find more sentences with positive text, not utilizing self-reference, utilizing invalidation. In paper [20], author proposed that when user used slang word consequently then it's less frequent and easily find the result by using text analysis. In paper [21], author proposed that teenager spend lot of time in social site and adult spend less time. These social networks turns into the significant methods for communicating their sentiments to the world. In paper [22], author proposed that adult users share their photos, videos or links in the different page. From this photos, videos or links give us the information that was initiated in the tweet.

So, Text classification is a crucial issue to be solved over the world. There are few research studies on the text classification and there are many ambiguities in it. This finding is important because we can classify the age from the text.

2.3.1 Existing Research Work Regarding Text analysis for Other Text

Some of research work on text analysis has been done for the age predicting by using both NLP and machine learning techniques. In paper [12] the researchers used three machine learning techniques-SVM, MaxEnt and Naïve Bayes to classify the text of Twitter messages with emoticons. They apply emoticons in the training corpus and train the corpus by using these three techniques. In the paper [13] four approaches were proposed (Topical approach, Emotional approach, Retrieval approach, and Lexicon approach) to calculate the emotional score of English text in messenger logs for six individual classes. And compare to other procedures Topical approach gives the best performance. In the research paper [14] the author proposed a supervised text classification framework and used data with specific features like hashtags and emoticons from twitter as sentiment labels to train the sentiment classifier using K Nearest Neighbors (KNN) algorithm.

In paper [15] [16] [17] the authors used LDA (Latent Dirichlet Allocation) model to extract emotion from the particular text document. In paper [15] the author proposed a joint emotion-topic model by augmenting LDA with an additional layer for emotion modeling and experiment it by using online news collection show. In paper [16] the researchers introduced an LDA based model for interpreting sentiment, and they used it for giving rank to the tweets concerning their popularity. In paper [17] both LDA and SVM have been used to specify the opinions from IMDB movie review dataset.

2.4 Conclusion

In this book, we introduced a text analysis system for predicting the age by using AI and machine learning techniques. This system can deal with different types of algorithm and special python programming function to remove the special character and unexpected data that are not useful for our project.

Chapter 3

Proposed Model

3.1 Introduction

To make an automated system for Age classification here, the author utilized AI systems because of its demonstrated precision level. Among numerous classes of AI arrangement moves toward the managed learning approach has been utilized on account of its similarity in displaying and controlling powerful frameworks. In supervised learning techniques, there need to provide a training dataset in which the algorithms are applied to trained the data. In addition, Logistic regression using bag of word and TF-IDF both are used for comparative analysis. Afterward, by using those trained data algorithms try to predict the text for any new post from social media.

3.2 Proposed Models

Here, this section represents the proposed model for classifying different age group which is consider two phase the data extract from different blog and then the classification phase.

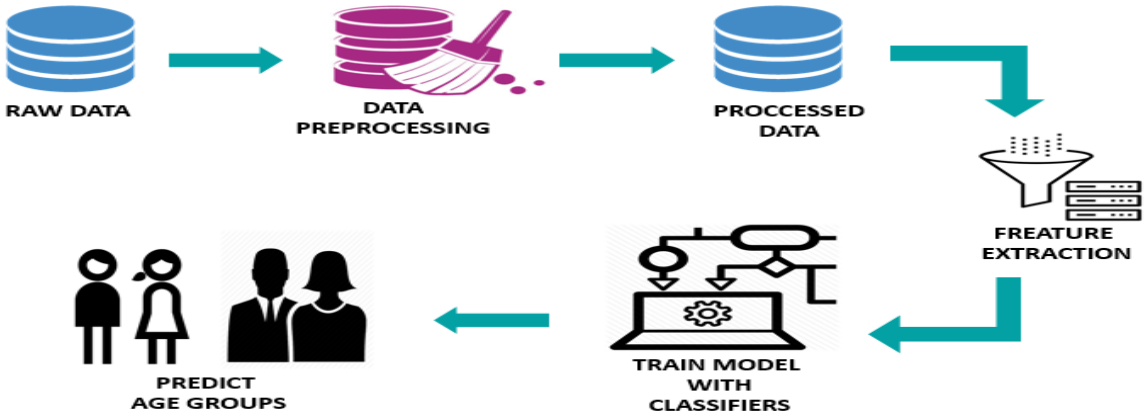

Fig. 3.1: Architectural overview of proposed model

At first, we take a blog dataset [23] for our model. Then we preprocessed this dataset by using different types of processes such as convert xml to csv, remove stop words, numbers, special characters, html tags, use lemmatizing, steaming, standardization etc.

After that, we used bag of words and TF-IDF for feature extraction. In this way, the program of feature extraction checks whether the word is accessible in the lexicon or not. By then, in the event that the word is accessible, at that point program adds that word to the recommendation list. In any case, on the off chance that the word isn't accessible in the lexicon, at that point the program checks its information words and makes a connection. As a result of relationship, on the off chance that the score esteem is more noteworthy than or equivalent to 65 percent, at that point that word from the information lexicon affixes to the proposal list, else, it goes for the accompanying word for making a correlation from the information word reference. Subsequent to completing the computations with all words from the information lexicon, it checks the recommendations list whether it is vacant or not. In the event that it is vacant, at that point it keeps a similar word as the client has given in the info information. In actuality, on the off chance that the proposal list isn't vacant, at that point it finds the most extreme score from the recommended words. On the off chance that similar scores show up for numerous words, at that point it adjusts the word by utilizing unigram technique. Else, it essentially chooses the word having greatest score worth and afterward it joins the word to the yield string. At that point it goes for the following info word. At the point when the estimation is done for all information words then the joined string will be supplanted as anticipated right sentence. We get two types of dictionary one for Bag of word and another for TF-IDF. Then we use different type of algorithm such as Random Forest, Multinomial Naïve Bayes, Linear SVC, and Logistic Regression with that two types of dictionary. From all that algorithm, we found accuracy, precision, recall, F1- score for our model. Finally, If we input a text post from our model’s input then it will predict the age group of post’s owner post.

3.3 Conclusion

To provide the results of any systems with good accuracy a well-organized procedure model is a very important factor. Through this model whole working procedure can visualize in an instance. We provide here a simple overview of our transliteration system for English language

Chapter 4

Data Preparation

4.1 Introduction

To lead the procedure with better performance first and most important part is to set up a reasonable information. Here we get Blogs data-set from web [23], where number of elements of data-set has around 0.6 million. Complete data-set split into two sections one was preparing training another was test data-set. Utilized 70 percent data for making preparing data-set and another 30-percent information for making test data-set.

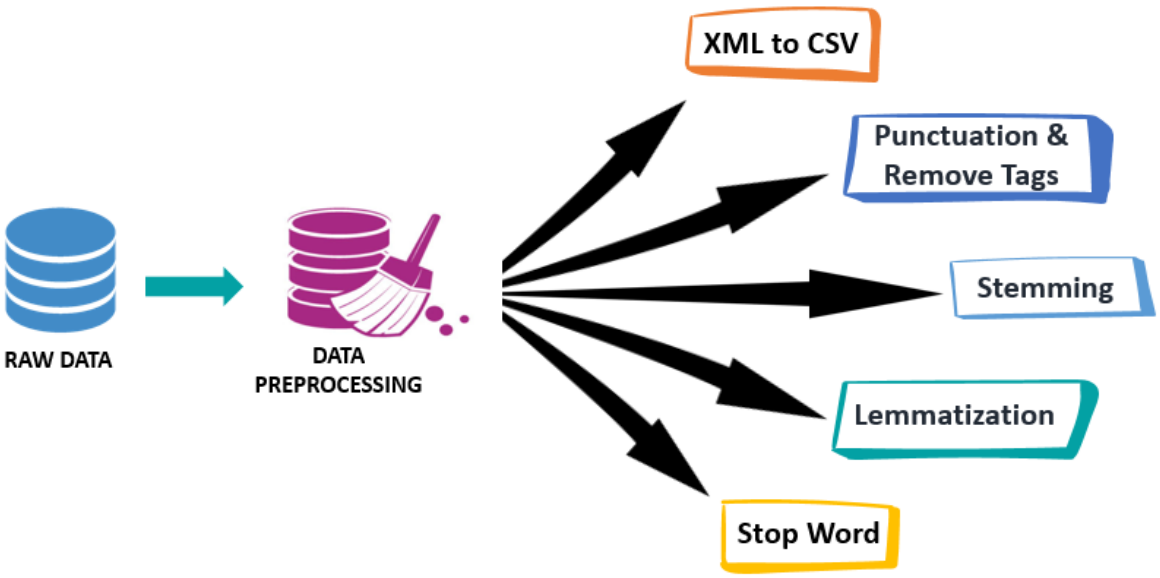

Fig. 4.1: Dataset Preparation

4.2 Training Dataset Preparation

We used 80 presents of total data for creating training dataset. To make it compatible for the execution, we manipulate this dataset based on the needs. We used to remove special characters, punctuation mark etc. We also use eliminating tags, standardization, lemmatizing, steaming, tokenization, stop word, NLP, Parts of speech predicting, text modeling etc. for preprocessing the training dataset.

4.3 Lexicon Preparation

Our lexicon divided into two dictionaries. One is for only the unique value another one is for all words. We need both of the dictionaries, the first one is for the correct the error word the second one is for the finding the probability when two or more word appears with the same Levenshtein Distance. Punctuation mark like " ।", "?", "!", ",", ":", ";" then the system omits these marks for better calculation

4.4 Standardization

Standardization means remove numeric value from data. By “review = re.sub(r"\b\d+\b",' ', review)”, we standardize our data. For example,

Table 4.1: Standardization

Before standardization	After standardization
AB12C	ABC
100 DCD 200	DCD
ABCDE 006Z	ABCDE Z

4.5 Eliminating Tags

By this process, we remove all HTML tags from our dataset. When we use html, we use some different tags for paragraph, link, head, table etc. By pat_tags = re.compile(r'<.*?/?>') and review = re.sub(pat_tags, ' ', review) we remove this type of HTML tags from data. Because tag are not necessary for data. Tags are one kinds of noise in data.

4.6 Natural Language Processing (NLP)

Natural Language Processing (NLP) is widely described that, it is the language of the natural language that can be automatically manipulated like speech and text, by using of the software. From the beginning of PC uses, programmers have been attempting to compose program that can comprehend language, for example, English. Basically PC doesn't comprehend the language in the manner that by and large people do. But now-a-days they can do a lot. The procedure of perusing and comprehend the language is extremely unpredictable for a computer. Natural Language Processing (NLP) has done it in efficient way where it likewise manages content analysis, data mining, spell correction, subtract undesirable image and word, create spam classifier and machine interpretation. So the Natural Language Processing can be encouraged our PC to comprehend as language utilized by humans. Twitter supposition examination can be actualized by utilizing information mining strategy with Natural language processing. To make arrangement with NLP needs some handling apparatuses to comprehend the language. Natural Language Tool kit(nltk) is open source library which is created in python to apply NLP techniques. These devices are containing vital devices to content procedure and slant investigation.

4.7 Tokenization

Tokenization is called tokenizer or lexer. It is a program that can take an array that is called string of character and for tokenization need to split it.

At Natural Language processing pipeline, the first step is sentence Tokenization and second step is word Tokenization. In the first step, the record will part with discrete sentence. Then it will be simpler to a developer to compose a program that can be comprehend a solitary sentence at that point comprehend the entire sentence. In the second step, the sentence break into independent words which is classified "token" and the procedure is called Tokenization.

We can better understand if we see an example:

Input: Everyone needs to care himself

Output: “Everyone”, “need”, “to”, “care”, “himself”

It is one of the easiest ways to for tokenization in English. It will split the word when it find the space between the words.

4.8 Stemming

Stemming is a significant idea of Natural language Processing that comes while extricating a few highlights out of negligible sentence or corpus of a great deal of sentence. The procedure of stemming in any sentence the inflected word is decreasing to their base or root structure. Assume there are some word which happens inside different sentence in a record, for example, "Intelligently", "Intelligence", "Intelligent" the base type of that word is "Intelligent". Another way "goes", "going", "gone" is changed over into "go". This is called stemming. So, we are not copying different words when we have a similar word in different forms. We are not accepting those words as different words as opposed to we are accepting those word as same word. It sets aside less effort to investigation a sentence or a document. Stemming is utilized where significance isn't significant, for example, spam discovery in a content.

We can import library from nltk by utilizing

#import nltk

#from nltk.stem import PorterStemmer

4.9 Lemmatization

In past area which is looking at stemming we can see a base or root structure is "Intelligent" that doesn't make any meaning. So, we can figure out that halfway portrayal of the word might not have any meaning. When working in a PC with text, it realizes the root type of the word which clarify that the different sentence are clarifying a similar idea.

This is the algorithmic procedure that is the purpose of the lemma word. This lemmatization depends on their necessary proposed work. Not at all like stemming, lemmatization relies upon effectively distinguishing the proposed grammatical form and importance of a word in a sentence, just as inside the bigger setting encompassing that sentence.

So, the lemmatization is same as stemming yet middle of the road portrayal of word in base structure which has contained some meaning. If you are doing some sort of examination where significance is significant then you can utilize lemmatization. It sets aside more effort to investigation and this procedure is utilizing where significance of a word is significant, for example, question noting applications.

4.10 Stop Word

A stop word is a common word (a, an, the, be, to, and) where search engine create a program for ignoring the common word. Sometimes this type of word doesn't need to our project work. So, it is very important to ignore this common word for our better result. When a programmer creates a program, they try to remove the unnecessary word by using the search engine. Sometimes programmer remove the certain list of word that are not needed to our project work is called stop list.

So stop words have no importance since they are not ready to express any unique significance dependent on some specific context.so when we are doing assumption investigation these words have no effect on opinion whether the slant is sure or negative. This is the explanation in the majority of the cases we truly need to expel these different stop words to show signs of improvement performance.

Fundamentally prevent words are identified from the rundown of realized stop words and there is no standard rundown of stop words which is reasonable for all application. It can change after relying upon your application.

4.11 Test Data Preparation

We choose some random sentence for making test dataset. Those sentences taken from different blog, social media post and comment. To test the accuracy and capability of this text classification system we prepare test dataset. Here we use 30 percent of total data as test dataset. That's 30 present random data from use for test our model and get the accuracy as a result. We try to use 2000 letters text of a entity from dataset.

4.12 Conclusion

A proper dataset is one of the crucial things for any machine learning techniques. To create this dataset manually is a big task to do. Mining the data in a proper way is a big challenge. Make usability for our task fully automatically is quite difficult for some garbage values.

Chapter 5

Methodology

5.1 Introduction

To conduct our text analysis an automated system for Age classification here, the author utilized AI methods because of its demonstrated precision level. Among numerous classifications of AI grouping approaches the machine learning approach has been utilized due to its similarity in displaying and directing powerful frameworks. In supervised learning techniques, there need to provide a training dataset in which the algorithms are applied to trained the data. In addition, Logistic regression using bag of word and TF-IDF both are used for comparative analysis. Afterward, by using those trained data algorithms try to predict the text for any new sentences or articles. To dealing with age classification from text, we proposed four approaches of supervised learning techniques. Logistic regression, Multinomial Naive Bayes, Linear SVC, Random Forest. We select these approaches for their better performance. Our text classification system will try to classify any sentence for age detection.

5.2 Feature Extraction

In order to obtain exact prediction of age group different method are applied. Character N-gram, vectorizing (Bag of word), TF-IDF for extracting many parameter. Every word are considered for feature. All the unnecessary punctuation mark, comma, html tag and many same ending were removed from the sentences.

5.2.1 Bag of word

We chose a bag-of-words model as a feature of the classification that can reviews opinion of the reviewers that may be text title or review text tags and other tags contains data that is irrelevant data for rating. Bag of words widely used and simplest language model for natural language processing. This model will help to find frequency of a word exists in a sentence. Make a table from how many times word appears in a sentence and then doing normalize and make a single vector. The bag of Words (BoW) model is the path of extracting feature from the text data which use in machine learning algorithms in this approach; we use the tokenized words for each observation and find out the frequency of each token. Here the example of Bag of Word-

Document 1

The quick brown Fox jumped over The lazy dog's Back.

Document 2

Now is the time for all good men to come to the Aid of their party.

Term	Document 1	Document2
aid	0	1
all	0	1
Back	1	0
Brown	1	0
Come	0	1
Dog	1	0
Fox	1	0
Good	0	1
Jump	1	0
Lazy	1	0
men	0	1
New	0	1
Over	1	0
Party	0	1
Quick	1	0
Their	0	1
time	0	1

Stop word List
For
Is
of
The
to

Fig 5.1: Example of Bag of word

5.2.1.1 n-gram

Character n-gram is a contiguous sequence of n character extracted from a given word. We extract character n-grams of length one (unigram), two (bigram) and three (trigram), and use these as features of the classifiers. Compared to word n-grams, which only capture the identity of a word and its possible neighbors, character n-grams are additionally capable of detecting the morphological makeup of a word. It detects the patterns in such misspellings.

Machine learning algorithm typically engaged in text analysis approaches comprises the following features:

N-grams convey constant sequences of n items in the conferred text and extensively used methods for natural language modeling. Here the example-

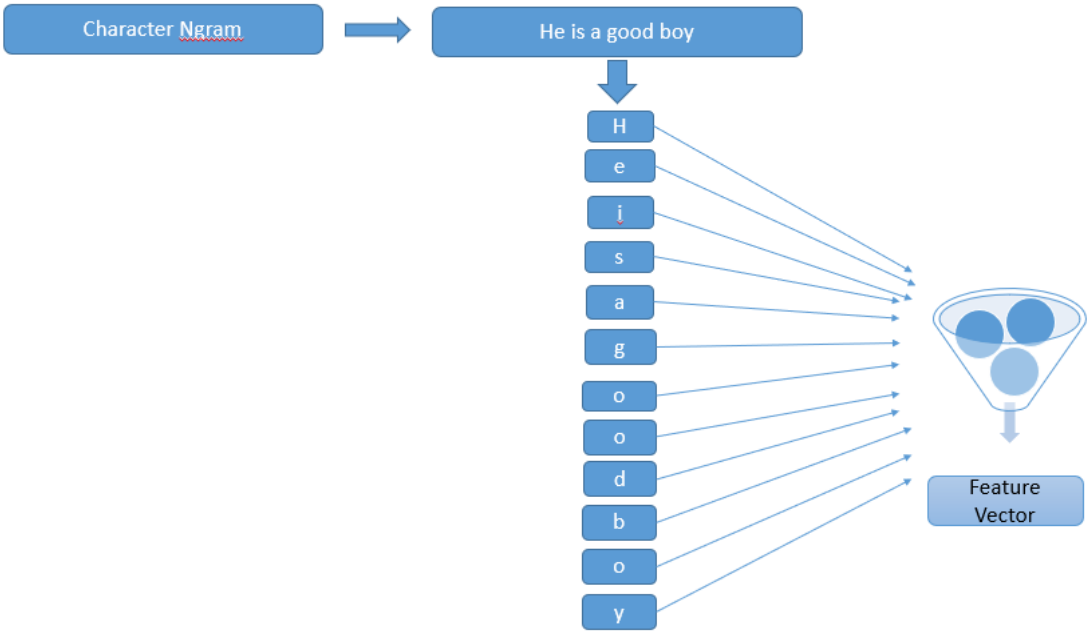

Figure 5.2: Uni-gram

5.2.1.2 Bigram

Bigram or diagram is a dilution of two contiguous items from the tokens that can be letters or words. The word n-grams of size two are called bigrams (n grams of length 2). For an example, suppose in a sentence, “I want to go there” here the bigrams are “I want”, “want to”, “to go”, and the trigrams are “I went to” and “went to cinema”. In n gram model it determines the instance of a word based on the instance of its (n - 1) which illustrate previous words. For bigram model it predicts the instance of a word given only its previous word.

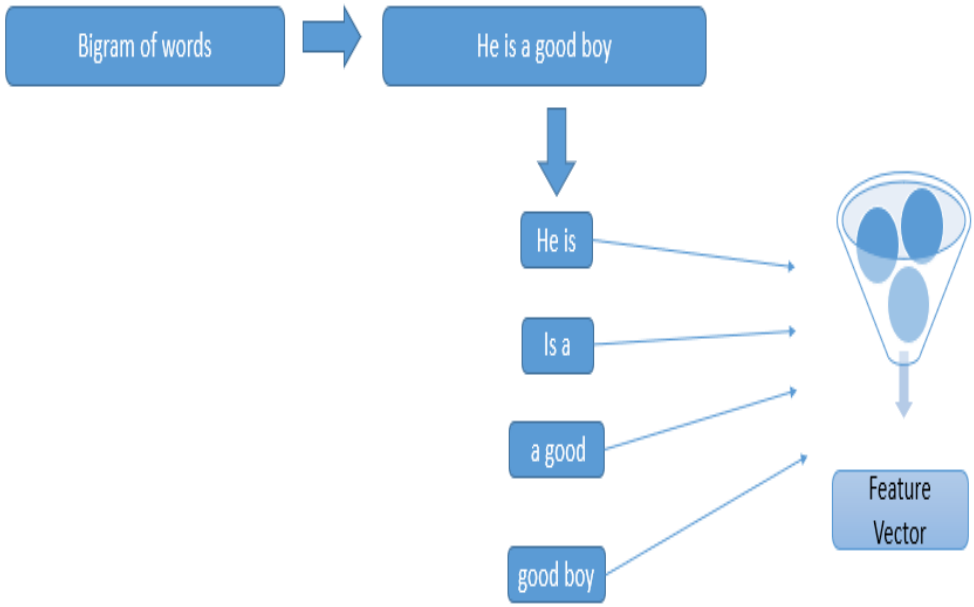

Figure 5.3: Bi-Gram

5.2.1.3 Trigram

The word n-grams of size three are called trigrams (n grams of length 3) and they are special case of n gram. For example, “I want to go there”, here the trigrams are “I want to” and “want to go.” In a trigram model it evaluate the instance of a word based on the previous words as n ? 1 = 2 for this case. Here the example-

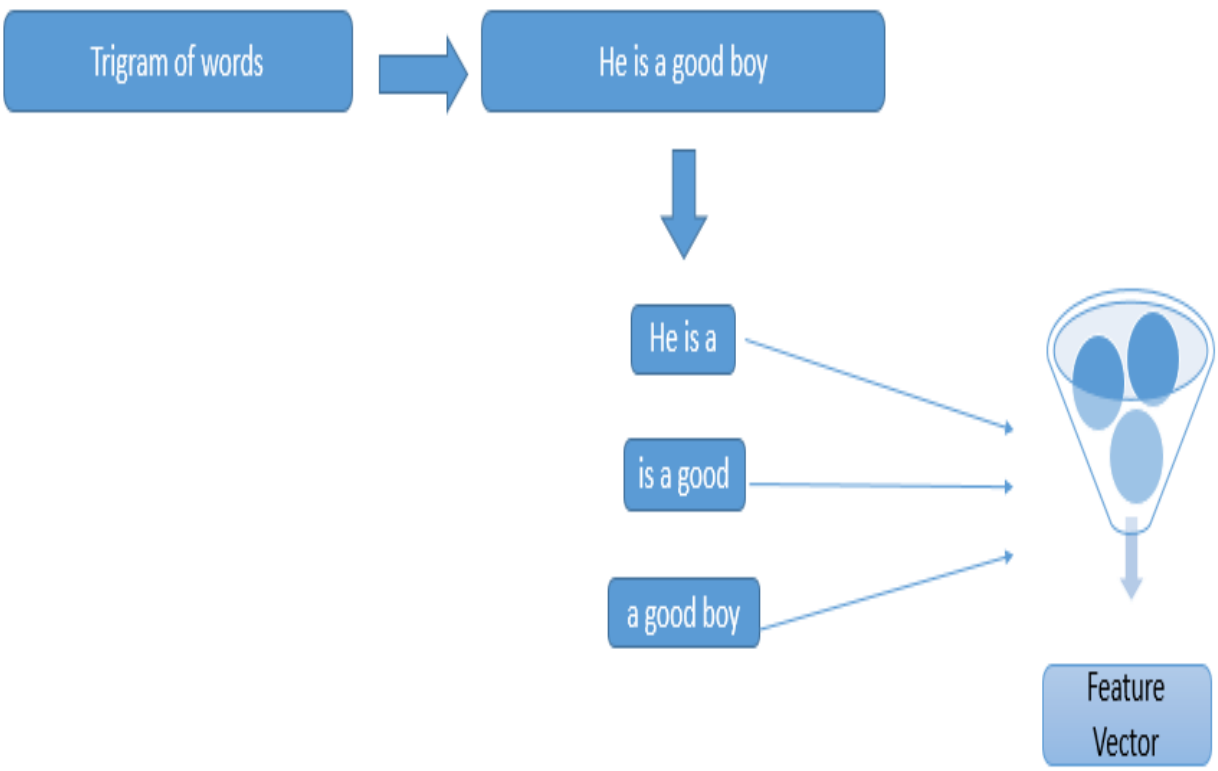

Figure 5.4: Tri-Gram

5.2.2 TF-IDF

In this approach we try to find word occurrence how many times on a sentence j In IDF we find number of sentence and divide with how much sentence have i.

$$TF(i,j)=\frac{Term\ i\ frequency\ in\ sentence\ j}{Total\ number\ of\ words\ in\ j}$$ ---------- 5.1

$$IDF(i)=\log \frac{Total\ documents}{Total\ Documents\ with\ term\ i}$$ ----------5.2

Multiply Tf score with Idf score we find score of a word in a document.

Score=TF(i,j)*IDF(i) ----------5.3

Example: Taking two sentences

• today is going to rain.

• i am not going today.

The frequency of most frequent word:

Table 5.1: The frequency of most frequent word

Frequency List	
Word	Frequency
today	2
is	1
going	2
to	1
rain	1
i	1
am	1

Now the Term Frequency for each of the different words in the histogram of frequency:

Table 5.2: Term Frequency for each of the different words in the histogram of frequency

Term Frequency		
Word	Sentence 1	Sentence 2
today	0.2	0.2
is	0.2	0
going	0.2	0.2
to	0.2	0
rain	0.2	0
i	0	0.2
am	0	0.2

According to formula of Inverse Document Frequency the IDF values are:

Table 5.3: Inverse Document Frequency

Inverse Document Frequency	
Word	IDF Values
today	0
is	0.69
going	0
to	0.69
rain	0.69
i	0.69
am	0.69

Now we have the values of TF & IDF model. Then calculating the final value ofTF-IDF model which is given below

Table 5.4: TF-IDF Values

TF-IDF Values							
words	going	is	to	today	i	am	rain
Sentence 1	0	0.14	0	0.14	0	O	0.14
Sentence 2	0	0	0	0	0.14	0-14	0

5.3 Classifier

Classifier in the machine learning is the supervised learning algorithm where the attribute is the nominal. Classifier is used after learning process to classify the data. In our model classifier used to classify the age group. Here we used different types of classifier algorithm such as Logistic Regression, Multinomial Naïve Bayes, Linear SVC and Random Forest.

5.3.1 Logistic Regression

Logistic regression is a statistical method which used for binary classification. It represent the result binary number 0 and 1. Here, it is used sigmoid function to convert word into numeric value and threshold value is 0.5. If the word value less than 0.5 then the word's binary value is 0. If the word value is greater than 0.5 then the word value is 1.

Logistic regression in Sentence level:

Step1:

Generate a hypothesis that can help us to classify age. We know from linear regression that

h_Θ (x)= $\Theta^T x$ ---------- 5.4

But we need a sigmoid function that can make our predicate value 0 to 1.Using this sigmoid function

g(z)= $\frac{1}{1+(e^{-z})}$ ---------- 5.5

we find equation h_Θ (x)= $\frac{1}{1+e^{-\Theta^T x}}$ ---------- 5.6

Step2: our data set has 10000 features Logistic regression hypothesis is the following:

h_Θ (x)=g($\Theta_0 + \Theta_1 x_1 + \Theta_2 x_2 + \cdots \Theta_{9999} x_{9999}$) ----------5.7

Step3:

If value is greater than 0.5 we can detect age is between 13-18

5.3.2 Multinomial Naïve Bayes

Step1: First need to find prior probabilities.

P(13-18)=Number of sentence classified in 13-18/Total number of sentence

Step2:

Next need to apply likelihood of a word exists in a sentence that will help to extract age of a person age by his text.

Step 3:

Need to apply another likelihood of a 13-18 on age occurring in a sentence followed that 13-18 goes to a sentence.

5.3.3 Random Forest

We have Bag of words from our sentences. Consider all of them as a feature which features help to make decision that a sentence is written by teenage or not .But in Random Forest feature/words has been taken randomly and make multiple decision tree from that we will get vote to predict a sentence.

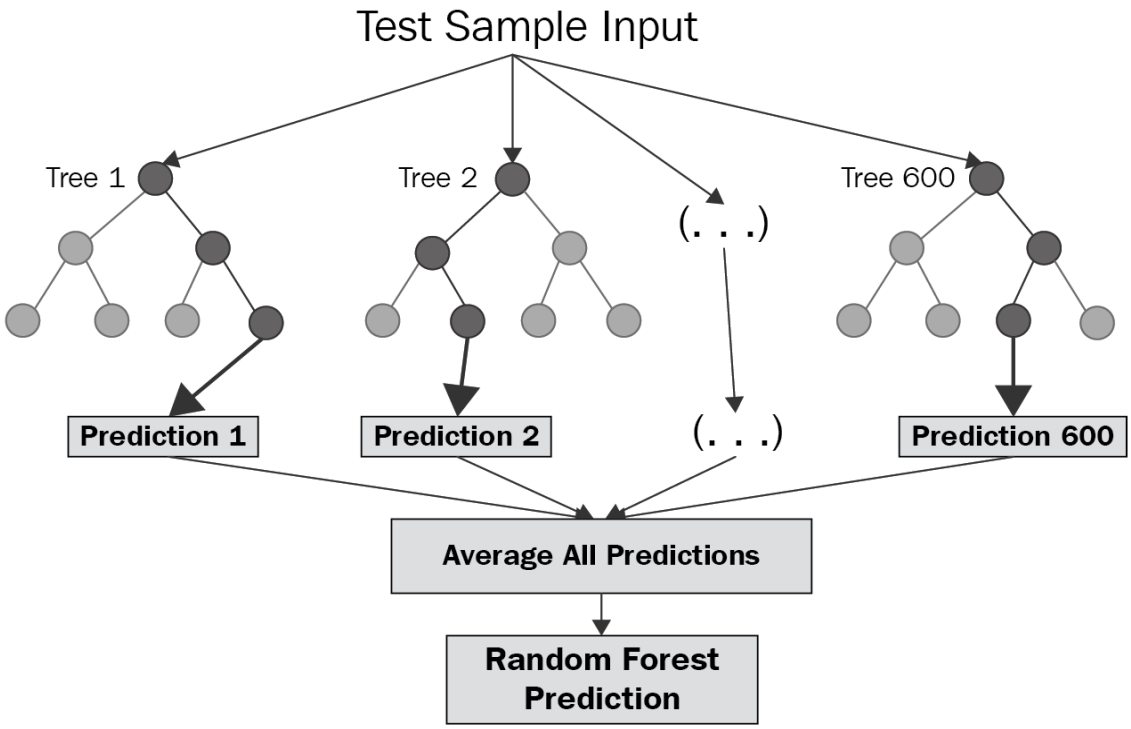

Figure 5.5: Random Forest

5.3.4 Linear SVC

We vectored words from documents. And those vectors create some points on 2D plane. Then need to find how remote vectors compare to origin.

X(x1, x2, x3,x4) where x1, x2, x3,x4 are vectored words .

Calculate length of vector $\|x\|=\sqrt{x_1^2 + x_2^2 + x_3^2 + x_4^2}$ ---------- 5.8

Direction of vector $\{\frac{x_1}{||x||}, \frac{x_2}{||x||}, \frac{x_3}{||x||}, \frac{x_4}{||x||}\}$ ---------- 5.9

Find the relation between vectors u.v=|u||v|COS(θ) ---------- 5.10

After finding relation need to create a line which divides two classes.

Hyperplane described as mx+c=0 as it is a linear line

mx-c=1 means inside or with the boundary it defines a class and mx-c=-1 inside or with the boundary it defines another class.

5.4 Data Processing

To build up this structure various data are expected to make a data lexicon for foreseeing or suggesting any word. These data are gathered from different online Blog. But those data are not pure data as some exceptional character (like:" : ; \/@ # $ % and ? | ! + - _, etc) can be worked up there. Along these lines, data filtering is expected to remove pure information from debased or garbage information. At the time of data filtering any sort of sections, , special characters are expelled. For example:

Table 5.5: Data processing

Sentence With Garbage	Removing Special character	Data Dictionary
http:// OK! Hello good people>>/I am not going	OK Hello good people I am not going	Good go Hello I people OK

After getting pure text data in a word format, the next step is to consider every word as an attribute then we apply many classification algorithms for predicting age group.

5.5 Conclusion

We have used four popular algorithm Logistic Regression (Bag of word,TF-IDF). In our program, we have used these two algorithms in two level of scope, one is sentence level, and another one is article level. As we collect data from different blog so having all the test data in the dataset is quite tricky. To solve this problem, we use a binning technique which gives a value to the missing data.

Chapter 6

Result and Analysis

6.1 Introduction

In this paper, to evaluate the performance of the proposed methods four different Machine learning algorithms were performed for classification of two age groups. We measure the performance and we checked which algorithms works better for our data set. We have done the analysis based on highest precision, recall, F1 score and accuracy. From the analysis we illustrate that we get highest precision for Logistic Regression (TF-IDF).

6.2 Result Analysis

To suggest any word from the data dictionary the system finds out the percentage of the similarity between two words. For every Algorithm, we have decided the exhibition for sentence level of extension by utilizing both proposed strategies for some individual classes. 10000 English sentences were utilized as our test data. At that point we determined the exactness for both proposed techniques. Something very similar has been done to quantify the exhibition in article level of scope. Numerous articles from various online news entry (Blog) have been utilized as our test data. As stated, before we use 10000 sentences for classification. 70000 Sentences valid for this age classification. From this 70% sentences were used for training the model and 30% percent were used to evaluate the performance of the model.

In the training phase logistic regression (TF-IDF) obtain a precision of 0.82 for teenager (13-18) and .78 for adults (26+) whereas the recall values .77 for teenager and .81 for adults (26+).

From this data we calculate the F- measure which is the harmonic average between information obtain. In this research the F-measure is 0.80. From the table as can be seen all the values.

6.2.1 Logistic Regression with Bag of Word

Table 6.1: Logistic Regression with Bag of word analysis

Age Group	Precision	Recall	F1 Score	Accuracy
Teenage	0.81	0.77	0.79	79.037%
Adult	0.77	0.81	0.79	79.037%

6.2.2 Logistic Regression with TF-IDF

Table 6.2: Logistic Regression with TF-IDF analysis

Age Group	Precision	Recall	F1 Score	Accuracy
Teenage	0.83	0.79	0.81	80.627%
Adult	0.79	0.82	0.81	80.627%

6.2.3 Multinomial NB with Bag of Word

Table 6.3: Multinomial NB with Bag of word analysis

Age Group	Precision	Recall	F1 Score	Accuracy
Teenage	0.79	0.76	0.78	77.429%
Adult	0.79	0.76	0.78	77.429%

6.2.4 Multinomial NB with TF-IDF

Table 6.4: multinomial NB with TF-IDF analysis

Age Group	Precision	Recall	F1 Score	Accuracy
Teenage	0.79	0.78	0.78	77.429%
Adult	0.77	0.78	0.77	77.429%

6.2.5 Random Forest with Bag of Word

Table 6.5: Random Forest with Bag of word analysis

Age Group	Precision	Recall	F1 Score	Accuracy
Teenage	0.78	0.77	0.78	78.224%
Adult	0.79	0.79	0.79	78.224%

6.2.6 Random Forest with TF-IDF

Table 6.6: Random Forest with TF-IDF analysis

Age Group	Precision	Recall	F1 Score	Accuracy
Teenage	0.79	0.79	0.79	78.245%
Adult	0.78	0.78	0.78	78.245%

6.2.7 Linear SVC with Bag of Word

Table 6.7: Linear SVC with Bag of word analysis

Age Group	Precision	Recall	F1 Score	Accuracy
Teenage	0.81	0.77	0.79	78.831%
Adult	0.77	0.81	0.79	78.831%

6.2.8 Linear SVC with TF-IDF

Table 6.8: Linear SVC with TF-IDF analysis

Age Group	Precision	Recall	F1 Score	Accuracy
Teenage	0.82	0.79	0.81	80.401%
Adult	0.79	0.82	0.80	80.401%

6.3 Comparative Analysis

Table 6.9: Comparative Analysis

classifier	Accuracy	precision	recall	f1 score
Logistic Regression (bag of word)	78.691	0.79	0.79	0.79
Logistic Regression (TF-IDF)	80.609	0.81	0.81	0.81
Multinomial NB (bag of word)	77.199	0.77	0.77	0.77
Multinomial NB (TF-IDF)	77.764	0.78	0.78	0.78
Random Forest (bag of word)	78.232	0.78	0.78	0.78
Random Forest (TF-IDF)	78.104	0.78	0.78	0.78
Linear SVC (bag of word)	78.432	0.79	0.79	0.78
Linear SVC (TF-IDF)	80.275	0.8	0.8	0.8

Here, we discuss the accuracy, precision, recall, F1- score for bag of word and TF-IDF feature extraction of our algorithm that are applied in our model

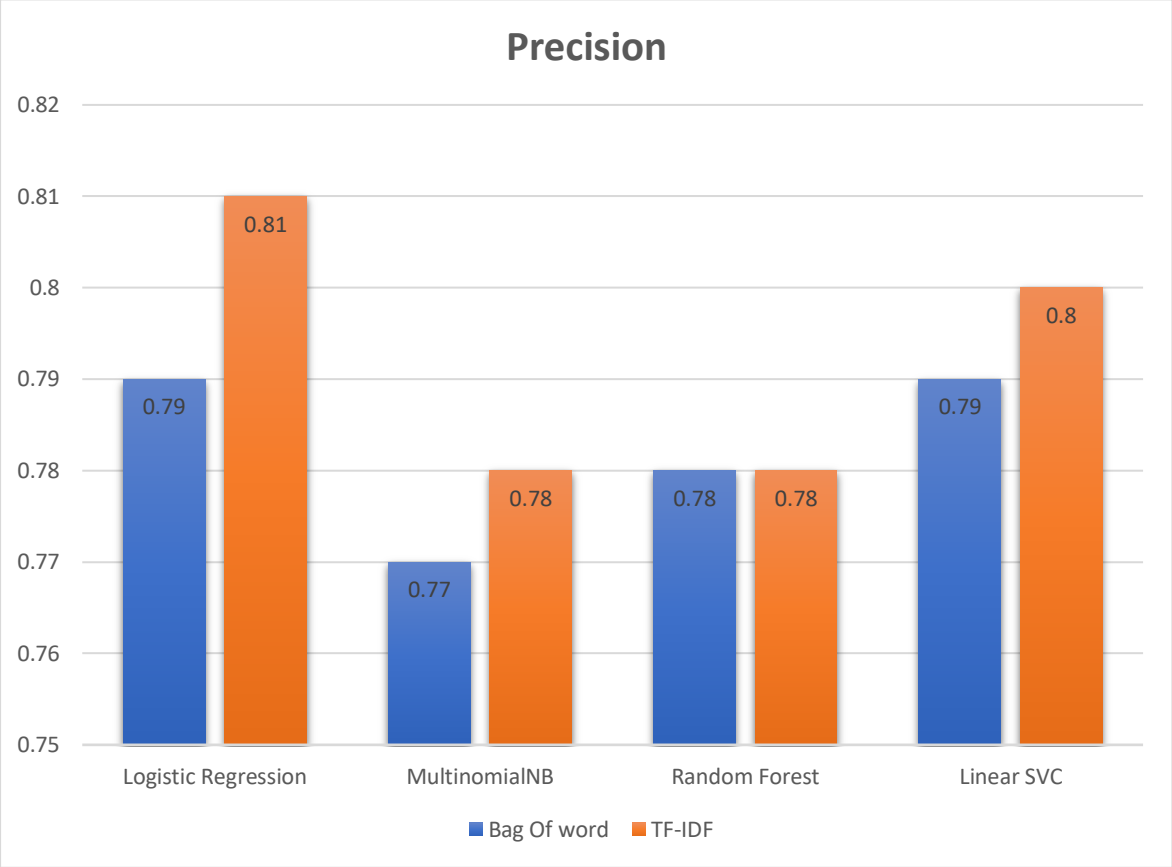

Figure 6.1: Precision vs classifier for bag of word and TF-IDF

In this graph, For Logistic Regression, precision for bag of word is 0.79 and precision for TF-IDF is 0.81. For Multinomial Naïve Bayes, precision for bag of word is 0.77 and precision for TF-IDF is 0.78. For Random Forest, precision for bag of word is 0.78 and for TF-IDF is 0.78. For Linear SVC, accuracy for bag of word is 0.79 and precision for TF-IDF is 0.80. Finally, we can see that precision is better for TF-IDF feature extraction than the bag of word extraction. Here, logistic regression provide the best performance to find the better precision.

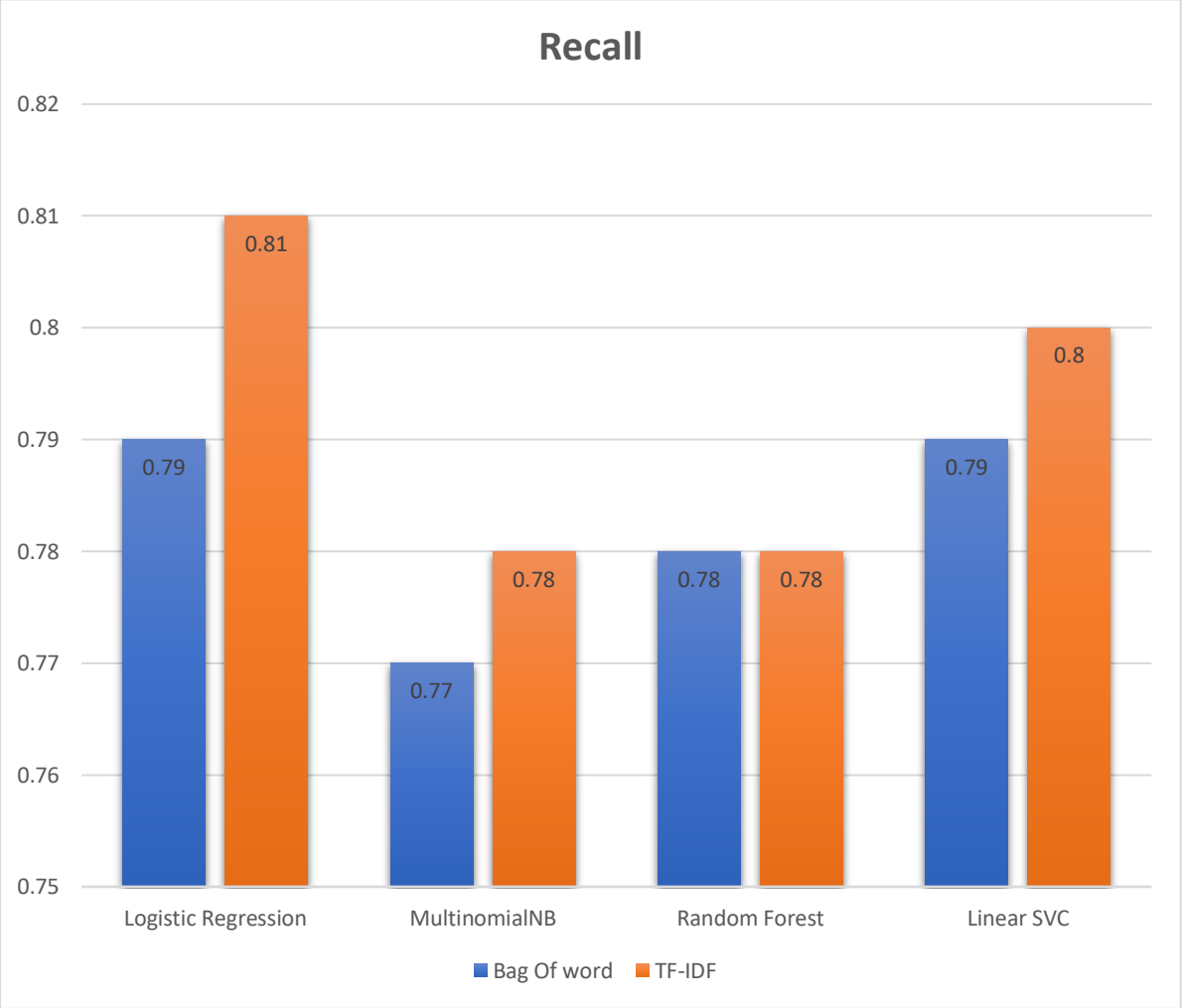

Figure 6.2: Recall vs classifier for bag of word and TF-IDF

In this graph, For Logistic Regression, recall for bag of word is 0.79 and recall for TF-IDF is 0.81. For Multinomial Naïve Bayes, recall for bag of word is 0.77 and recall for TF-IDF is 0.78. For Random Forest, recall for bag of word is 0.78 and recall for TF-IDF is 0.78. For Linear SVC, accuracy for bag of word is 0.79 and recall for TF-IDF is 0.80. Finally, we can see that recall is better for TF-IDF feature extraction than the bag of word extraction. Here, logistic regression provide the best performance to find the better recall.

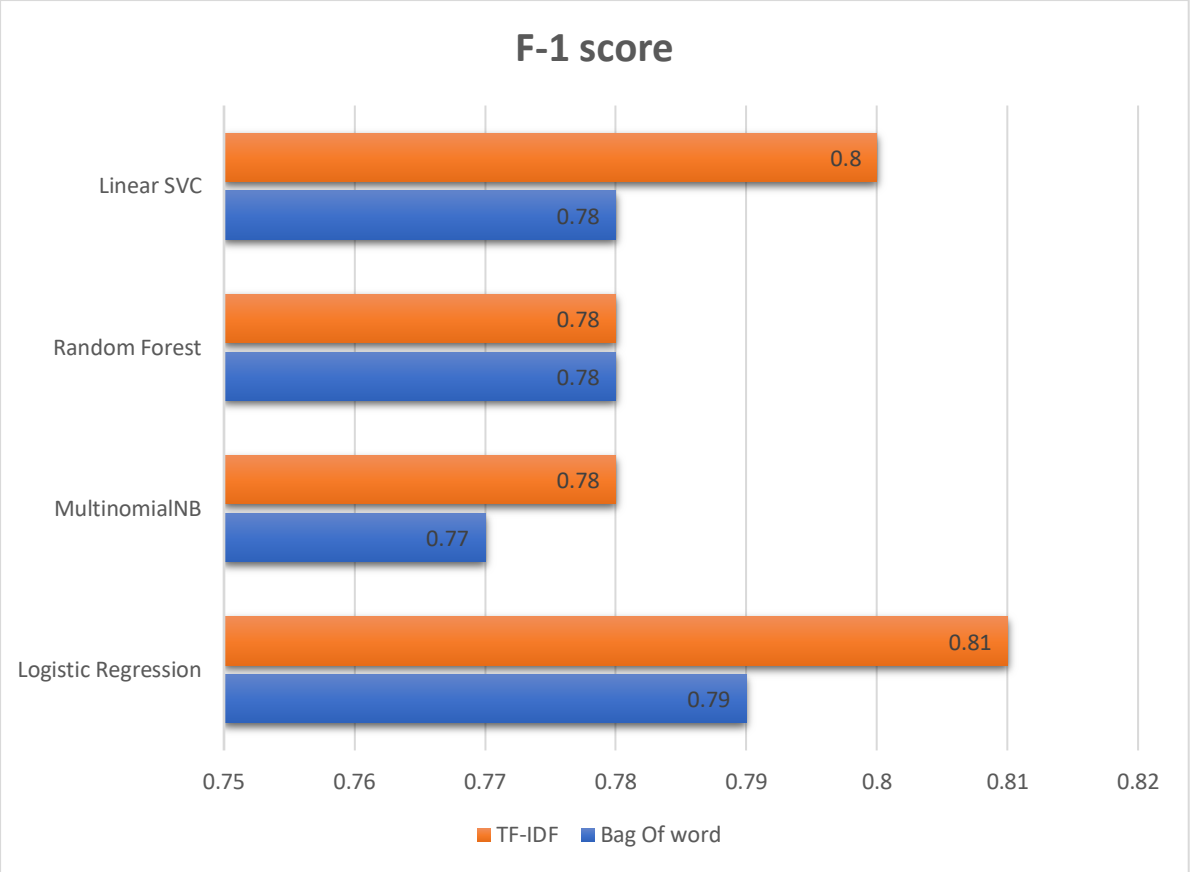

Figure 6.3: F-1 score vs classifier for bag of word and TF-IDF

In this graph, For Logistic Regression, F-1 score for bag of word is 0.79 and F-1 score for TF-IDF is 0.81. For Multinomial Naïve Bayes, F-1 score for bag of word is 0.77 and F-1 score for TF-IDF is 0.78. For Random Forest, F-1 score for bag of word is 0.78 and F-1 score for TF-IDF is 0.78. For Linear SVC, accuracy for bag of word is 0.78 and F-1 score for TF-IDF is 0.80. Finally, we can see that F-1 score is better for TF-IDF feature extraction than the bag of word extraction. Here, logistic regression provides the best performance to find the better F-1 score.

So, F1-score give the same value as like as precision and recall because

F1 score = $\frac{2*precision*recall}{precision+recall}$ ---------- 6.1

and the prediction of true positive, true negative, false positive, false negative are correctly classified and the model is correctly balanced.

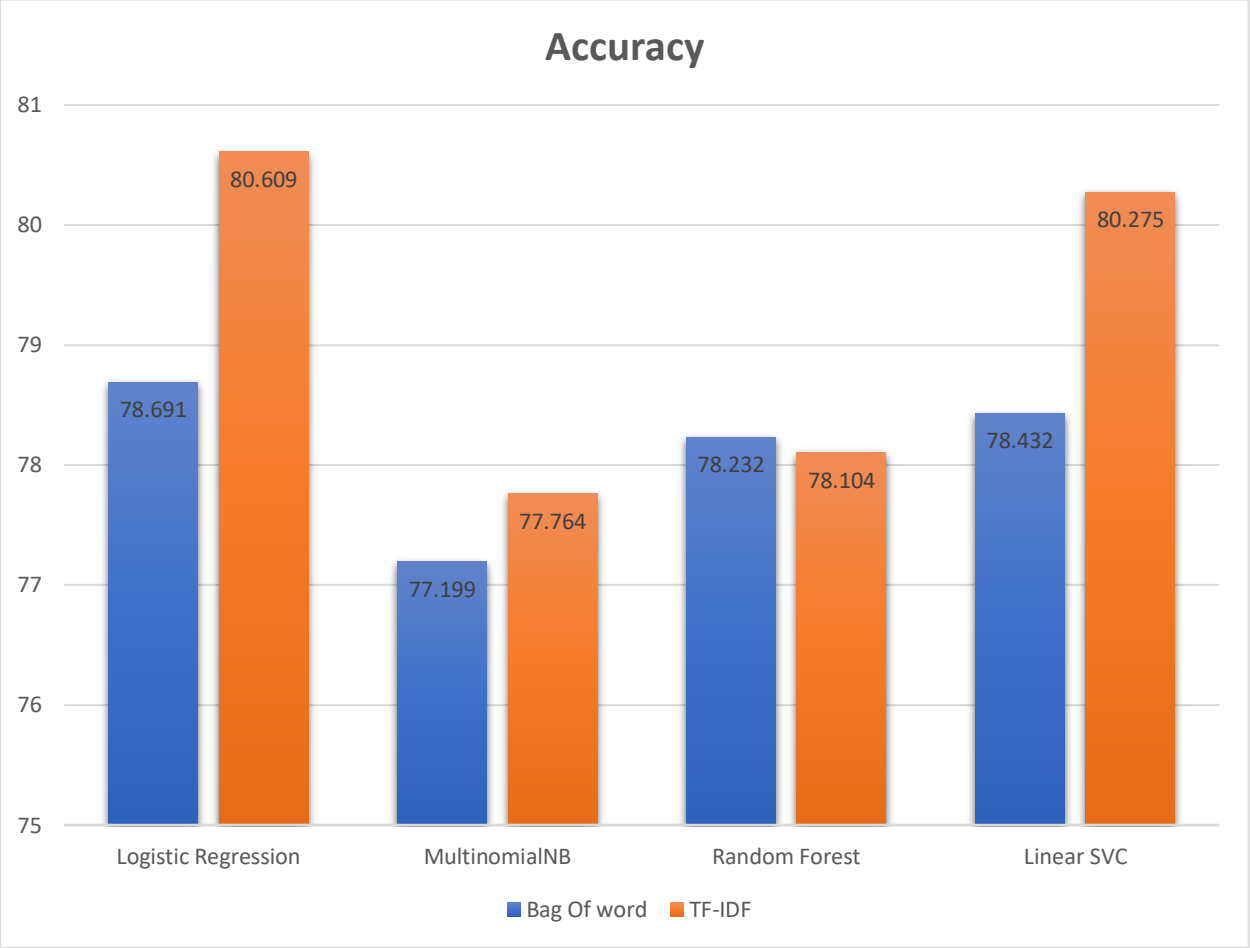

Figure 6.4: Accuracy vs classifier for bag of word and TF-IDF

In this graph, For Logistic Regression, accuracy for bag of word is 78.691 and accuracy for TF-IDF is 80.609. For Multinomial Naïve Bayes, accuracy for bag of word is 77.199 and accuracy for TF-IDF is 77.764. For Random Forest, accuracy for bag of word is 78.232 and accuracy for TF-IDF is 78.104. For Linear SVC, accuracy for bag of word is 78.432 and accuracy for TF-IDF is 80.275. Finally, we can see that accuracy is better for TF-IDF feature extraction than the bag of word extraction. Here, logistic regression provides the best performance to find the better accuracy.

We experimented our proposed method on mixed age dataset where the Logistic regression (TF-IDF) classifier perform better than others and give more accuracy which is79.394%. Even we can see its recall, f1-score are also better than others.

6.4 Conclusion

We implemented algorithms and features and can see the result of our prediction for the performance of our text classification. Both Logistic Regression TF-IDF and Linear SVC work well for sentence level of scope with the accuracy of 80.609% and 80.275% respectively

Chapter 7

Conclusion and Future Work

7.1 Conclusion

Based on today's perspective text has become a treasure trove of revealing useful information and people's opinions regarding anything. So uncover the views from the text is an important task now for so many fields like product analysis, social media monitoring, market research and analysis and so on. Based on these needs our paper works on detecting age by using five proposed methods. We achieved a satisfying accuracy of above 79.384% for Logistic regression TF-IDF on the word level. Whenever we type a sentence, for each word it calculates the age based on the text which slightly different. If an input word is not available in Lexicon, then the process will not work for it. Our proposed model will help to know the specific age groups of as given text form the bag of word. This model will be beneficial for social media monitoring, market analysis and so on. It will also open new doors for text mining, especially for English text analysis. With the proper utilization of this model, people will be beneficial, and new fields in text mining will be opening.

7.2 Future Work

For further work, this framework can be deployed in other languages and also can be implemented in any large platform including any language search engine by developing its performance. We improve our dataset and create our won algorithm and Improve our model for multi age group and gender.

References

[1] R. G. Guimarães, R. L. Rosa, D. De Gaetano, D. Z. Rodríguez and G. Bressan, "Age Groups Classification in Social Network Using Deep Learning," in *IEEE Access*, vol. 5, pp. 10805-10816, 2017.

[2] S.M.Sawyeretal.,‘‘Adolescence:Afoundationforfuturehealth,’’Lancet, vol. 379, no. 9826, pp. 1630–1640, Apr. 2012. [22] J. Y. Jang, K. Han, P. C. Shih, and D. Lee, ‘‘Generation like: Comparative characteristics in instagram,’’ in Proc. 33rd Annu. ACM Conf. Human Factors Comput. Syst., Seoul, South Korea, Apr. 2015, pp. 4039–4042

[3] J. Y. Jang, K. Han, P. C. Shih, and D. Lee, ‘‘Generation like: Comparative characteristics in instagram,’’ in Proc. 33rd Annu. ACM Conf. Human Factors Comput. Syst., Seoul, South Korea, Apr. 2015, pp. 4039–4042.

[4]. S. Utz and N. C. Krämer, ‘‘The privacy paradox on social network sites revisited: The role of individual characteristics and group norms,’’ J. Psychosoc. Res. Cyberspace, vol. 3, no. 2, pp. 73–79, Nov. 2009.

[5] Amitava Das , “Opinion Extraction And Summarization From Text Documents In Bengali” , A., & Bandyopadhyay, S. Phrase level polarity identification for Bengali, International Journal of Computational Linguistics and Applications, 1(2), pp. 169–181, 2010. [online] Amitavadas.com. Available: http://www.amitavadas.com/Pub/Amitava_Das_PHD_Thesis.pdf [Accessed 6 Oct. 2018].

[6] K. M. Azharul Hasan, Md Sajidul Islam, G. M. Mashrur-E-Elahi, Mohammad Navid Izhar, “Sentiment Recognition from Bangla Text”, Technical Challenges and Design Issues in Bangla Language Processing, 2013. Available:https://www.researchgate.net/publication/277792568_Sentiment_Recognition_from_Bangla_Text.

[7] C. Peersman, W. Daelemans, and L. Van Vaerenbergh, ‘‘Predicting age and gender in online social networks,’’ in Proc. 3rd Int. Workshop Search Mining User-Generated Contents, Glasgow, Scotland, Oct. 2011, pp. 37–44. [Online]. Available: http://doi.acm.org/ 10.1145/2065023.2065035

[8] L. Zheng, K. Yang, Y. Yu, and P. Jin, ‘‘Predicting age range of users over microblogdataset,’’Int.J. DatabaseTheoryAppl.,vol.6,no.6,pp.85–94, Oct. 2013.

[9] D.-P. Nguyen et al., ‘‘Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment,’’ in Proc. 25th Int. Conf. Comput. Linguistics, Dublin, Ireland, Aug. 2014, pp. 1950–1961.

[10] D. A. Huffaker and S. L. Calvert, ‘‘Gender, identity, and language use in teenage blogs,’’ J. Comput.-Mediated Commun., vol. 10, no. 2, pp. 1–24, Jun. 2005.

[11] J. W. Pennebaker and L. D. Stone, ‘‘Words of wisdom: Language use over the life span,’’ J. Personality Social Psychol., vol. 85, no. 2, p. 291, Aug. 2003.

[12] Asif Hassan, Mohammad Rashedul Amin, Abul Kalam Al Azad, Nabeel Mohammed, “Sentiment Analysis on Bangla and Romanized Bangla Text (BRBT) using Deep Recurrent models”, Available: https://arxiv.org/ftp/arxiv/papers/1610/1610.00369.pdf

[13] Shaika Chowdhury, Wasifa Chowdhury, “Performing sentiment analysis in Bangla microblog posts”, 2014 International Conference on Informatics, Electronics & Vision (ICIEV), 23-24 May 2014. Available: https://ieeexplore.ieee.org/document/6850712

[14] Alec Go, Richa Bhayani, Lei Huang, “Twitter Sentiment Classification using Distant Supervision”.Available:https://cs.stanford.edu/people/alecmgo/papers/TwitterDistantSup ervision09.pdf

[15] Dmitry Davidov, Oren Tsur, Ari Rappoport, “Enhanced Sentiment Learning Using Twitter Hashtags and Smileys”, Coling 2010: Poster Volume, pages 241–249, Beijing, August 2010. Available: http://www.aclweb.org/anthology/C10-2028

[16] Suvarna D.Tembhurnikar, Nitin N.Patil, “Sentiment Analysis using LDA on Product Reviews: A Survey”, International Journal of Computer Applications (0975 – 8887) National Conference on Advances in Communication and Computing (NCACC 2015). Available: https://pdfs.semanticscholar.org/2482/900623fcec6d42459dbbe2b2d4119e76ee14.pdf

[17] Raja Mohana S.P, Umamaheswari K, Ph.D, Karthiga R, “Sentiment Classification based on Latent Dirichlet Allocation”, International Journal of Computer Applications (0975 – 8887) International

Conference on Innovations in Computing Techniques (ICICT 2015). Available: https://pdfs.semanticscholar.org/983a/5f269a1333379d81b7ebb043bf57b2da7247.pdf

[18] Hasan, Md. "Sentiment Analysis with NLP on Twitter Data." PhD diss., East West University, 2018.

[19] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker, ‘‘Effects of age and gender on blogging,’’ in Proc. AAAI Spring Symp., Comput. Approaches Anal. Weblogs, Stanford, CA, USA, Mar. 2006, pp. 199–205.

[20] H. A. Schwartz et al., ‘‘Personality, gender, and age in the language of social media: The open-vocabulary approach,’’ PLoS ONE, vol. 8, no. 9, pp. 73–79, Nov. 2013.

[21] L. A. S. Shapiro and G. Margolin, ‘‘Growing up wired: Social networking sites and adolescent psychosocial development,’’ Clin. Child Family Psychol. Rev., vol. 17, no. 1, pp. 1–18, Mar. 2014.

[22] T. A. Pempek, Y. A. Yermolayeva, and S. L. Calvert, ‘‘College students’ social networking experiences on Facebook,’’ J. Appl. Develop. Psychol., vol. 30, no. 3, pp. 227–238, Jan. 2009.

[23]http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm?fbclid=IwAR1unSPLHEMgvd0Y9OvbIWa0pPBook2nod5Wrchm-JJzaby2rWyaHq9pQ1s