


Flesch-Kincaid Measure as Proxy of Socio-Economic Status on Twitter: Comparing US Senator Writing to Internet Users

Samara Ahmed, King Abdulaziz University, Saudi Arabia

Adil Rajput, Effat University, Saudi Arabia*

 <https://orcid.org/0000-0002-0310-7175>

Akila Sarirete, Effat University, Saudi Arabia

Tauseef J. Chowdhry, NUML University, Pakistan

ABSTRACT

Social media gives researchers an invaluable opportunity to gain insight into different facets of human life. Researchers put a great emphasis on categorizing the socioeconomic status (SES) of individuals to help predict various findings of interest. Forum uses, hashtags, and chatrooms are common tools of conversations grouping. Crowdsourcing involves gathering intelligence to group online user community based on common interest. This paper provides a mechanism to look at writings on social media and group them based on their academic background. The authors analyzed online forum posts from various geographical regions in the US and characterized the readability scores of users. Specifically, they collected 10,000 tweets from the members of US Senate and computed the Flesch-Kincaid readability score. Comparing the Senators' tweets to the ones from average internet users, they note 1) US Senators' readability based on their tweets rate is much higher, and 2) immense difference among average citizen's score compared to those of US Senators is attributed to the wide spectrum of academic attainment.

KEYWORDS

Big Data, Natural Language Processing, Social Computing, Social Media, Socioeconomic Status (SES)

1. INTRODUCTION

1.1. Motivation and Background

Social computing garnered significant attention after the advent of Web 2.0. The extensive use of blogs, Myspace communities, and various online forums affected the way people conducted social interactions (Parameswaran & Whinston, 2017). Social media platforms offer a unique chance to perform social science and online research. It offered users a forum to voice their views unequivocally since they don't need to reveal their true identity. While many social media platforms today require the end-users to confirm their real identity, the process is not always perfect. In addition, federal regulations bind social media companies to protect the real identity of the end-user. The 2004 US

DOI: 10.4018/IJSWIS.297037

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License
Copyright (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and reproduction in any medium, provided the author of the original work and original publication source are properly credited.

presidential campaign, for example, popularized the idea of online advertising and encouraged many scholars to research its influence (Weinberg & William, 2006).

The launch of Amazon Mechanical Turk in 2005 brought a new dimension to the area of Artificial Intelligence (Irani, 2017). The crowdsourcing platform allowed the users to outsource tasks to humans, which would be difficult for a computer to perform. The crowdsourcing platform allows advertisement of a task for a group of users who will perform it for an incentive (money, contribution to literature, etc.). The social media platform has the concept of crowdsourcing embedded in it, as pointed out by (Paniagua & Korzynski, 2017). As an example, Twitter was used successfully in various domains such as emergencies; disaster relief, etc. in the context of crowdsourcing (Jordan et al., 2018) - discussed more in the next section. In these scenarios, the experts depended on the feedback from volunteers in the affected region, based on which agencies could come up with an appropriate real-time response. Such scenarios come under the umbrella of active crowdsourcing. Passive crowdsourcing, on the other hand, involves soliciting user action without the users consciously realizing that they are contributing. The concept of hashtag on twitter where various users would contribute to a particular topic is one example of passive crowdsourcing. In this scenario, people interested in soliciting feedback can start a hashtag that can help gather valuable information.

Social and medical sciences researchers have begun to focus on the vast number of available data. Although social network data are not the means by which a particular individual's problems are identified or treated by themselves, the data can be used to identify different symptoms as measures for certain problems of certain issues in mental health (Rajput & Ahmed, 2018a). The techniques developed in the field of Natural Language Processing (NLP) can be invaluable in the processing and segmentation of text information, as needed by social and medical science practitioners, using the various segmenting techniques. The choice of the corpus is one of the main requirements to these steps. We use the definition of the corpus as "a collection of naturally occurring text, chosen to characterize a state or variety of a language" (Schvaneveldt et. al., 1976). In general, constructing a corpus includes considering a specific text to the problem and deriving keywords, bigrams and sometimes trigrams (two or three-word sentences) that are used excessively in a given area. As an example, (Rajput & Ahmed, 2018b) argue that a corpus should be developed to assist mental health professionals in detecting depression among users provided some group of people. The researchers base their observations on the twitter hashtag # depression. The study gathered overwhelmingly evident terms and found that these words are part of the language of depression patients. Once such a corpus is established, researchers would look at a random text and predict with a certain assurance whether the words used by the individual are the same frequency as those in the corpus.

1.2. Problem Description

One of the factors that mental health practitioners and sociologists look at is the socio-economic status (SES) of a given individual. SES status helps in predicting the potential issues the individual might face. (Collins, 2016) discusses, for example, SES comorbidity and alcoholism. The level of education is one of the

main determinants of a person's SES. The education level, in turn, is correlated with the ability to write (Kellogg & Raulerson, 2007). (Geiser & Studley, 2002) in their findings argue that a student's ability to compose extended text is the single best predictor of success in finishing freshmen level coursework. (Flesch et al., 1975) defined the Flesch-Kincaid test by providing a formula to combine the proportions of a text with the grade level of the text, as follows.

1. The ratio of total words to total sentences in a given text
2. The ratio of total syllables to total words in a given text.

The grade level inferred from the above equation is directly linked to the school grade level of the given text. A score of 12, for instance, shows that a grade 12 student can understand the given

text while a score of 14 means that the text is written at a level of a second year University student. Specifically, the formula is described as follows¹:

$$\text{F-K readability} = 206.835 - 1.015 \times (\text{words/sentences}) - 84.6 \times (\text{syllables/words})$$

Given the above, our current work aims at establishing a measure that will serve as a proxy for SES from publicly available data. Deriving our motivation from the concept of crowdsourcing in social media, we explore the possibility of looking at the grade level of a writer to predict the education level of a person that in turn would indicate one aspect of his/her SES. The study was based on a group of people, all of whom were in possession of a college degree as a minimum requirement. A Pew Center research report Pew (2019) mentioned that every member of the US Senate in 2019 has attained at least a college degree. This is compared to only 34% of American adults aged 25 or older accomplishing the same feat. Thus, our study comprised of the following: 1) analyzing US Senators' impromptu writing and 2) comparing random online forum posts that reflect writing of an average member of American society to the above.

To this end, we looked at a public online discussion forum from two different sources

- a. Forum where users who bought a particular car posted their thoughts and impressions. The discussion forum had already been segmented into various regions in the USA and Canada (e.g., Northeast, South, Western Ontario, etc.). We scraped the forum and analyzed the data to see whether individuals from different regions differed in Flesch-Kincaid grade level.
- b. Randomly selected 5 communities on reddit²

The work contribution is summarized as follows:

1. Establish a baseline for US Senators' tweets (Assumption: they are representative of something being jotted down without much or very little preparation and are a reflection of writers' immediate thoughts)
2. Compare the Flesch-Kincaid grade level score obtained from US Senators to grade score level of user community from different regions of the US and Canada including reddit communities.

2. LITERATURE REVIEW

2.1. Relationship Between Reading and Socioeconomic Status (SES)

In the earlier work described in (Rajput et. al., 2019), the authors established from literature the three factors contributing to the SES status of a person namely education, income and occupation. The authors provided examples from literature discussing the effect of all the three variables on various facets of individuals' lives. In this paper, we focus on the academic aspect and specifically the work done in the realm of readability.

Earlier work in this area focused on establishing metrics that would measure reading and writing abilities of people in various income brackets. Income on the other hand, was easier to categorize based on multitude of factors such as self-reported data, housing prices, etc. (Chall & Jacobs, 1983) selected children - belonging to economically challenged families - from different grades and chose various metrics to measure progress in terms of reading and writing. Specifically, they looked at the skill of evolving writing ability from merely listing contents to story telling. The authors showed that children from grades five going on to grades seven started lagging behind in this area. The authors continued their work in this area and published their findings in their seminal work (Chall and Jacobs, 2003). The children's SES status was established by their eligibility for free-lunch program. The authors established five stages for reading from stage 0 to 5 where stage 0 referred to pre-reading and stage 5 referred to most mature reading stage. The authors argues that students transition from stage

0 to 5 by going through “learning to read” stage – characterized by stages 1 and 2. These stages are typically acquired in grades 1-3 while the next phase extends for a longer period of time where the students “read to learn” – stages 3 and 4. The authors study focused on students in grades 4 when they transitioned between the phases of “learning to read” to “reading to learn”. The authors noted that students coming from economically challenged background not only struggled from transitioning to fourth grade (reflective of their lagging behind in “learning to read” stage), same students continued to struggle in grades 7, 9 and 11.

(Bowe, 1995) presented similar results where the author looked at five-year olds’ preschool phonological development and the first grade reading skills’ development based on their paternal occupation. The author established that even after accounting for IQ status, the children belonging to lower SES (based on paternal occupation struggled in reading).

As opposed to the aforementioned work focused on English language (US and Australia), authors in (Heppt, et. al., 2015) looked at a subset of German students belonging to low SES native speakers and nonnative speakers. The results showed that the students struggled in acquiring basic reading skills necessary to learn and communicate their achievements.

Lastly, the work done by (D’angiulli et. al., 2005) focused on fifth graders in British Columbia, Canada and showed that the fifth grade children belonging to lower income levels needed remedial work during school years to come up to par to their counterparts in terms of reading skills.

Having established the importance of reading, researchers showed the correlation between poor reading and writing skills as discussed in the next subsection. Our work in this paper builds upon the findings that lower reading and writing skills usually correspond to poor performance in school and in turn reflects in the occupation/income status of the individual. We focus on gleaning the reading/writing skills of an individual from social media platform and public repositories. We argue that social media platform provides us ample evidence that can act as a proxy of an individual’s SES status by looking at the readability scores of their writings.

2.2. Measuring Readability

Having established the correlation of reading skills to SES, we will explore in this section the correlation of poor writing to both reading and in turn SES. Nevertheless, writing can occur under different conditions. It can be written in the form of a manuscript, impromptu or written extemporaneous (Blankenship, 1974). The manuscript form of writing follows a thorough process of thoughts, reflection, writing, and revision. The Impromptu writing process does not assume any prior deliberation while the extemporaneous form assumes a short time of reflection before writing the thoughts on paper (Cronn-Mills & Croucher, 2001).

Writing has received much attention as many scholars consider it an afterthought of the thinking process (Applebee, 1984; Emig, 1977; and Odell, 1980). Work done by (Howard, 1988) considers the act of writing as the father of thought. Across the literature, scholars agree that the language used in writing is superior to the oral form (Devito, 1965). (Chafe and Tannen, 1987) presented a detailed review of the work done on written and oral language. (Nippold et. al., 2014) divided the communication into three forms namely social, academic and practical and explored the use of complex syntax in conversational and narrative speaking. The results showed that people use complex syntax and sophisticated language when narrating a particular event as opposed to engaging in conversational mode.

The work done in (Blankenship, 1974) performed a detailed study on various forms of written and oral mode of communication and she divided the communication into a conversation, oral impromptu, written impromptu, oral extemporaneous, written extemporaneous and manuscript. The author employed various metrics such as sentence and word length, cloze score, adjective/verb quotient, and preposition/token quotient among others. The work is important to our study as we look at both twitter messages and floor speeches delivered by members of the US senate. The underlying assumption for

our work is that Twitter provides a mode for written impromptu/written extemporaneous, while the speeches delivered on the Senate floor reflect the manuscript mode.

One frequent method used by the United States Military explained in (Kincaid et al., 1975) - also adopted in word processing software – known as the Flesch–Kincaid Readability Tests (Stockmeyer, 2009). The Flesch Reading Ease Test rates a text on a 100-point scale with higher scores indicating easier readability (Flesch, 1948). The Flesch-Kincaid Grade Level Test standardizes the score to the U.S. graduate school levels (Kincaid et al., 1975). A score of 12, for instance, shows that a grade 12 student can understand the given text. Some researchers have begun to use such measures in order to start adopting to the realm of big data (Flaounas et al., 2013). Much research is being conducted on readability with some measures superior to the Flesch-Kincaid test (Si & Callan, 2001). Given important progress, however, there remain some key problems in terms of consistency – outlined in (Mailloux et al., 1995); (Wang et al., 2013). Because of the common use, we opted for the Flesch-Kincaid test as the base of our research. Although various methods exist in the literature, the Flesch-Kincaid test uses the length of the sentence and the words in the sentences, to measure the individual's level of education.

As discussed later in the paper, we establish a base case by gathering both the tweets by the members of the US Senate and their floor speeches. We chose the tweets randomly without paying attention to the context, content, and length of the speeches as it is beyond the scope of this work.

2.3. Crowdsourcing

Meriam-Webster³ defines crowdsourcing as “the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community rather than from traditional employees or suppliers”. (Doan et. al., 2011) first discussed the use of world-wide-web as a medium for crowdsourcing and mentioned four significant challenges namely: Way of recruiting contributors, gauging their abilities, combining the work performed and avoiding abuse of the system. The paper also discussed challenges in maintaining the quality of the work performed.

Crowdsourcing techniques, on the other hand, are used in various scenarios to help collect information quickly about large groups of people in social networks. (Wazn, 2017) reviewed crowdsourcing taxonomies, crowdsourcing research, regulatory and ethical aspects, including some prominent examples of crowdsourcing. The author concluded that crowdsourcing has the potential to be extremely promising, in particular in the healthcare, as it has the ability to collect information quickly and in a cost-effective way. In order to group users online into communities, researchers have made use of hashtags that will identify the interest of a community of users. It is worth noting that the crowdsourcing concept predated social media as many researchers worked on amalgamating data from heterogeneous sources for many years. One such approach is defined in (Adali et. al., 1995). This approach also became popular in the P2P paradigm (Rajput & Rotenstreich, 2004).

In this paper, we propose to group users on the basis of their geographical context and create a corpus for these users. We start by crowdsourcing a group of users that have achieved a high academic accomplishment and study the level of their writing in the case of impromptu writing. Furthermore, we use the terms impromptu writing/extemporaneous writing as synonyms despite the work by (Blankenship, 1974) as discussed above. The reason for this is that we have no mechanism that can help us decide whether the writing on twitter is an example of impromptu or extemporaneous writing. Note that extemporaneous writing assumes a short amount of deliberation while impromptu writing is a direct reflection of instantaneous thoughts of the writer as pointed out by (Blankenship, 1974). Once we have established a baseline, we will look at discussion forums for a particular commercial vehicle. We started exploring this in our earlier work (Rajput et al., 2019).

2.4. SES Application in Medicine and Psychiatry

The work of (Kawachi, 1999) built on the concept of social capital and reported that individual level factors - such as low income, low education, etc. - are strongly correlated with self-reported poor health. Social Capital is defined as “an individual’s personal network and elite institutional affiliations” (Belliveau et. al., 1996). (Veenstra, 2000) built upon the above where the author looked at the three elements of social capital namely trust, commitment, and identity and showed that both income and education were related to self-reported health data. The commencement of the century saw researchers study the comorbidity of low SES status with different diseases across the vast realms of medicine. Given the fact that our research can have a strong application in the mental health realm, we will briefly discuss various efforts in this area only. Full discussion of various areas that can be affected by SES is beyond the scope of this paper.

(Baker & Wagner, 1966) made the case that researchers and practitioners are ignoring the social aspect of patients when defining treatments and that the seeking of psychotherapy for children is inversely proportional to the social class of the patient and the caretakers. (Dohrenwend, 1990) observed that the 1980s established the relationship between SES status and various psychiatric disorders such as schizophrenia. In other words, high poverty levels were shown to be related to high levels of psychiatric disorders. However, the author argues that it had been difficult to unlock the riddle that would establish low SES as a cause or a consequence of psychiatric disorders. (Vitaro and Tremblay, 1999) showed that impulsivity in gambling had a high prevalence of low SES adolescent males. (Piko & Fitzpatrick, 2001) found a correlation between SES status and psychosocial health among Hungarian adults. (Mayes & Calhoun, 2011) studied the effect of various variables including SES on autistic symptoms and established that they had a higher rate of presence in lower SES. (Goldberg et. al., 2011) also established a relationship between lower SES and both schizophrenia and cognitive ability. The relationship between risk of hospitalization for schizophrenia, SES, and cognitive functioning was established in (Goldberg et. al., 2011). (Hanscombe et. al., 2012) studied the Gene-Environment interaction among a group of kids in the UK. (Bates et. al., 2016) further looked in this area and found no evidence that the SES status could alter the intelligence of an individual. Rather, the effect is confined to the development of abilities of the individual in various disciplines. In their paper, (Fernald et. al., 2013) discussed in their paper that SES differences might strongly result in clear differences in language processing and vocabulary development starting at 18 months. This is also the basis of our work, which is founded on the premise that the difference in SES will directly reflect on the way a person writes. Looking at a person’s writing across various media will provide a preponderance of evidence on their SES status. Such information can prove invaluable in studying the behavior of people across different SES and in turn help in detecting certain mental illnesses such as depression, as shown in (Rajput and Ahmed, 2018b).

3. METHODS

3.1. Participants

Our current work used NLP techniques to cull together data from the social media to focus on two specific groups

1. The US Senators and their public twitter handle to gather individual tweets
2. Average Americans spread over geographic locations in USA and Canada

3.2. Design

Preprocessing and Processing Data

We followed the following process for preprocessing and processing of data:

1. Used the Twitter API to gather 10,000 tweets from the US Senators' public Twitter account
2. Computed the individual Flesch-Kincaid score for each tweet
3. Computed the average score for 10,000 tweets for each senator
4. Computed the average of average Flesch Kincaid score for Senators to establish the baseline for impromptu writing for this group
5. Used the built-in 'urllib' functionality of python to handle all URL functionality
6. Used the beautifulsoup package to scrape the data that is present in the forum
7. Locate forums that are segregated by geographical regions in the USA and Canada
8. Anonymized the user data to ensure that the privacy of individuals is not violated (even though the data is public)
9. Computed the Flesch-Kincaid grade level for each text within a forum
10. Computed the average of each region and stored it in the database
11. Implemented the above on a standard Dell running Ubuntu Linux and Python3 program with a 16G RAM. Given that we did not have any performance requirements, the program can be ported to any platform that supports Python3. Furthermore, the software used in this project is freely available.

Challenges

One of the biggest challenges when gathering data is ensuring the legality of using the data – discussed in (Youyou et. al., 2015), (Ahmed & Rajput, 2020) and (Ahmed, 2019). All the data that we gathered is available from public sources and we use the NLP techniques and tools described in (Rajput, 2020). Specifically, we gathered the following set of data:

1. Tweets from the official twitter account of US Senators
2. Vehicle forum for a specific manufacturer divided by various regions. Such forums help the manufacturer learn from customers various issues and problems they might face. We chose the specific vehicle because it is in the choice of an average American. Specifically, we tied the average American income to the class of cars⁴ and looked for the cars in a particular category.
3. We gathered the posts in the aforementioned forum as a sample of impromptu writing

Furthermore, to ensure the anonymity of the users, we masked their online identity by assigning each user fictitious pseudonym before storing the information in a database.

The APIs used

For this work, we used Python programming to gather and analyze the data. We used the following open-source APIs available for python programming language.

1. Twitter API: This requires registering with Twitter and creating a Twitter development account. The twitter library can be installed for Python that provides all the requisite APIs
2. Pandas: This is an open-source python library which allows data cleaning, preparation and fast analysis. The data can be easily imported into Excel.
3. NLTK: This is one of the most powerful NLP libraries that provide basic tools such as tokenization, stemming, lemmatization etc. Interested readers can refer to [50] for pertinent details.
4. Sklearn: This library helps in big data analysis such as classification, regression, clustering etc.

4. RESULTS AND DISCUSSION

4.1. Results

Figure 1 below presents the average Flesch-Kincaid score for the tweets (impromptu writing) of the US Senators.

Table 1. Averages calculation US senators

Item	Average
Total Number of Senators	99 ⁵
Tweets per Senator	10,000
Average Tweets Score	12.06139
Standard Deviation	1.022358

Table 1 summarizes the average tweets scores and the standard deviation.

Figure 1. Average score for tweets (US Senators)

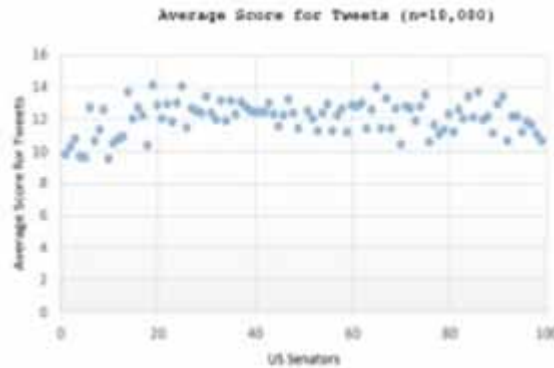


Table 2. Impromptu writing of an average American across the US

Region	Total Number of posts	Average Flesch Kincaid Score	Standard Deviation
Northeast	66	6.62	3.07
MidAtlantic	101	6.51	3.60
South	22	8.41	3.46
Midwest	92	6.70	3.44
Southwest	41	7.32	4.47
Northwest	12	7.50	2.71
West	36	5.58	2.47

A cursory look at the data above shows that the level of impromptu writings reflects the high academic attainment of the US Senators. Compare this to the data obtained from various forums that represent the impromptu writing of an average American across the continental United States presented in Rajput et. al. (2019) (See Table 2).

Looking at the above results, we see that average Flesch Kincaid score is significantly lower than that of US Senators who averaged more than 12. The problem with the above however, is that the number of posts are much less than the 10,000 tweets for the US Senators and hence might not be a suitable metric for comparison. Furthermore, the standard deviation for the posts is much higher than that of the one obtained for the tweets from US Senators.

To alleviate the above issue, we randomly selected 5 communities on reddit as follows:

1. r/Relationships
2. r/TodayILearned
3. r/Tinder
4. r/YouShouldKnow
5. r/GetMotivated

The rationale behind choosing the aforementioned communities were to ensure the diversity of thoughts and people who post to such communities. Please note the following:

1. As opposed to the auto forums used earlier, we could not separate the posts based on Geographical data
2. We ignored the images and the emojis that were posted
3. We randomly selected threads under each community and gleaned all the posts under each thread
4. The number of posts that we used were 100 to ensure that we can compare it to the results obtained by those from US Senators.

4.1.1. r/Relationships

The community at hand has various posts discussing issues regarding relationships. The following table presents the results.

Table 3. Averages calculation r/relationship community

Items	Average
Total Number of Discussion	100
Threads per Discussion	100
Total Posts	10,000
Average Post Score	7.4580
Standard Deviation	0.2927

The table above shows that the average Flesch-Kincaid score is around 7.5 which is much lower than those of the US Senators.

4.1.2. r/TodayILearned

Next, we take a look at the community r/TodayILearned. The results, while lower than those of US Senators, were slightly higher than the previous community.

Figure 2. Average Flesch-Kincaid score

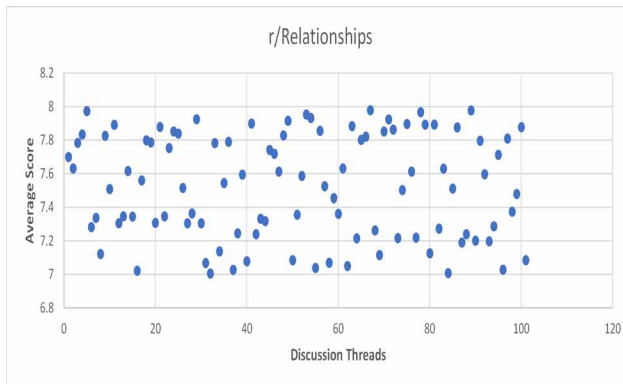
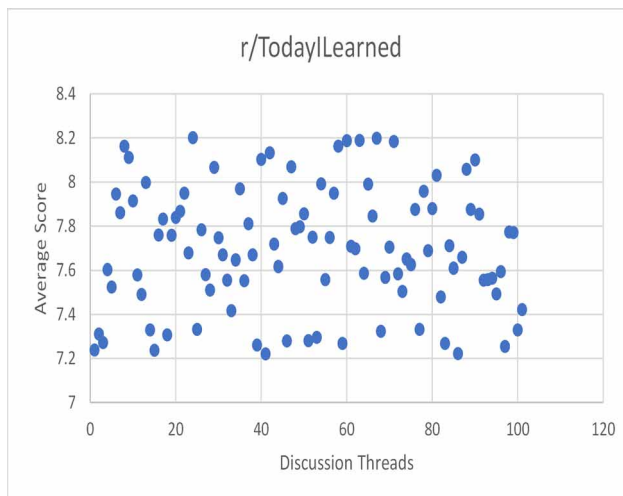


Table 4. Averages calculation r/todayilearned community

Items	Average
Total Number of Discussion	100
Threads per Discussion	100
Total Posts	10,000
Average Post Score	7.75
Standard Deviation	0.295

Figure 3. Average score for r/TodayILearned (Discussion threads)



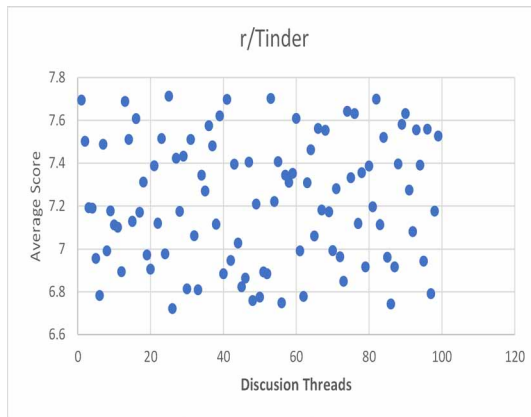
4.1.3. r/Tinder

Interestingly, the community r/Tinder while being very different than the earlier ones show strikingly similar results both compared to the other reddit communities and the posts by US Senators.

Table 5. Averages Calculation r/Tinder Community

Items	Average
Total Number of Discussion	100
Threads per Discussion	100
Total Posts	10,000
Average Post Score	7.200
Standard Deviation	0.290

Figure 4. Average score for Tinder (discussion threads)



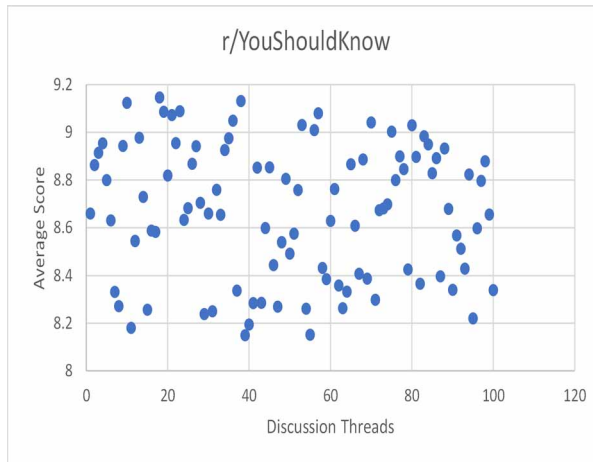
4.1.4. r/YouShouldKnow

The community r/YouShouldKnow displayed results similar to the other communities on reddit with the average Flesch-Kincaid result oscillating between 7 and 8.

Table 6. Averages Calculation r/Relationship Community

Items	Average
Total Number of Discussion	100
Threads per Discussion	100
Total Posts	10,000
Average Post Score	7.4580
Standard Deviation	0.2927

Figure 5. Average score for YouShouldKnow (discussion threads)



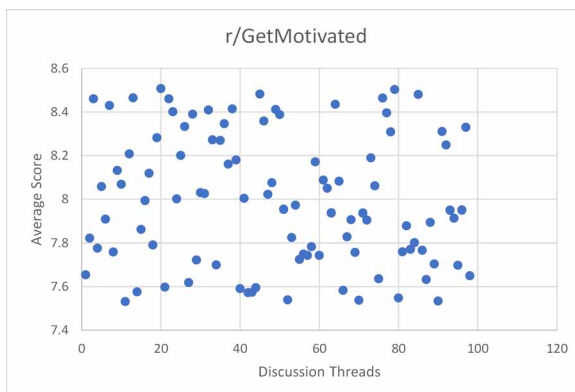
4.1.5. r/GetMotivated

Lastly, the r/getMotivated community writers displayed a slightly higher level than the remaining communities but still significantly lower than those of US Senators.

Table 7. Averages Calculation r/GetMotivated Community

Items	Average
Total Number of Discussion	100
Threads per Discussion	100
Total Posts	10,000
Average Post Score	8.003
Standard Deviation	0.3015

Figure 6. Average score for GetMotivated (discussion threads)



4.2. Discussion

The aforementioned results provide us a basis to compare and confirm various theories presented in the literature before the advent of Web 2.0 and the spread of social media.

To begin with, we confirm the findings of statistics reported by (Pew, 2019) which reported that only 34% of an average American adult aged 25 or older has a Bachelor's degree. Impromptu writing by US Senators had an average score of 12 compared to no more than eight of an average American citizen. Also, note that the data for average American was gleaned by looking at the average income based on each of the regions above and tying it to the types of vehicles they drive (Rajput et al., 2019). (Note that a difference of one on the Flesch Kincaid score is equivalent to a grade level.

Secondly, comparing the standard deviation of the Senators' tweets to an average American impromptu writing, the variation is much bigger for the general population. This can possibly be explained by many factors such as 1) Income disparity is much less among US Senators as opposed to an average American citizen and 2) one-third of average Americans have attained the same academic level as those of Senators. This group will score much higher and hence the high variation.

Thirdly, note that the corpus chosen for the average American is very specific and not from Twitter as we wanted to focus on responses for a very specific group – namely an average American with average income and a high probability that the person is a native English speaker. The average Flesch-Kincaid score consistently ranged between 6.5 and 8.3 but with higher standard deviation.

Fourthly, when we looked at the number of posts that we could confidently attribute to residents of USA and Canada, we felt that we needed a larger sample to confirm the findings. To alleviate this concern, we expanded our sample space by including five random communities with high number of posts. While we could no longer ascertain that the writers are from the USA and Canada, the posts were all in English. Looking across 10,000 posts for each of the five communities, we confirmed our earlier findings where all the communities – diverse in their topic of discussion – displayed an average Flesch Kincaid score between 7 and 7.5 with standard deviation of less than 0.5.

Fifthly, while many new readability measures have been proposed, Flesch-Kincaid remains the choice of many researchers till today. The measure is used in all aspects of science such as medicine and politics as displayed by the works in (Tahir et al., 2020), (Baier, S. 2021), (Somasundaram et al., 2021) and (Zhang et al., 2021) among others.

Lastly, one of the underlying assumptions of academic attainment is that the age of the person writing is 25 or older. While we can confirm this for the members of US Senate, we cannot state with absolute certainty that this is the case for the corpus we harvested. However, given that the forum is for owners, we have high confidence that the average age of writers on the forum is 25 or older as reported by CarMax⁶.

5. CONCLUSION

In this paper, we proposed a groundwork for predicting a proxy of SES of different communities. First, we established an analysis of the US senators' tweets and determined their Flesch-Kincaid average scores. We considered this as our baseline and we compared the results to different groups of online communities in US based on their geographic regions and analyzing their tweets for commercial vehicle selections. The results helped to note the difference in the choice of language and in turn act as an indicator of academic attainment – used as a proxy for SES status of the writer. Our work is comparing the impromptu writings for native speakers. In the future, we would like to focus on segregating native speakers' writings from non-native speakers. Furthermore, we currently assume Twitter to reflect a person's impromptu form of writing. We would focus on digging in the literature to devise algorithms that can help us accomplish this. Lastly, Our work focuses solely on writings in English language. We would like to replicate the study for another language and compare the results.

ACKNOWLEDGMENT

Please note that the work presented in this paper is an updated and revised version of the preprint described in (Ahmed et. al., 2020b).

REFERENCES

- Adali, S., Brink, A., Emery, R., Lu, J., Rajput, A., Rogers, T., . . . Ward, C. (1995). *HERMES: A Heterogeneous Reasoning and Mediator System*. <http://www.cs.umd.edu/projects/Hermes/publications/abstracts/hermes.html>
- Ahmed, S., Rajput, A. E., Sarirete, A., Aljaberi, A., Alghanem, O., & Alsheraigi, A. (2020a). Studying Unemployment Effects on Mental Health: Social Media versus the Traditional Approach. *Sustainability, 12*(19), 8130. doi:10.3390/su12198130
- Ahmed, S., Rajput, A. E., Sarirete, A., & Chawdhery, T. J. (2020b). *Social Media Platform: Measuring Readability and Socio-Economic Status*. Academic Press.
- Ahmed, S., & Rajput, A. (2020). Threats to Patients Privacy in Smart Healthcare Environment. In *Innovation in Health Informatics* (pp. 375-393). Academic Press. doi:10.1016/B978-0-12-819043-2.00016-2
- American Psychological Association. (2016). *Measuring socioeconomic status and subjective social status*. Public Interest Directorate, Socioeconomic Status Office, Resources and Publication.
- Applebee, A. N. (1984). Writing and reasoning. *Review of Educational Research, 54*(4), 577–596. doi:10.3102/00346543054004577
- Baker, E. H. (2014). Socioeconomic Status, Definition. *The Wiley Blackwell Encyclopedia of Health, Illness, Behavior, and Society*, 2210-2214. doi:10.1002/9781118410868.wbehibs395
- Baker, J. Q. (1966). Social class and treatment in a child psychiatry clinic. *Archives of General Psychiatry, 14*(2), 129–133. doi:10.1001/archpsyc.1966.01730080017003 PMID:5901392
- Bates, T. C., Hansell, N. K., Martin, N. G., & Wright, M. J. (2016). When does socioeconomic status (SES) moderate the heritability of IQ? No evidence for g× SES interaction for IQ in a representative sample of 1176 Australian adolescent twin pairs. *Intelligence, 56*, 10–15. doi:10.1016/j.intell.2016.02.003
- Belliveau, M. A., O'Reilly, C. A. III, & Wade, J. B. (1996). Social capital at the top: Effects of social similarity and status on CEO compensation. *Academy of Management Journal, 39*(6), 1568–1593.
- Blankenship, J. (1974). The influence of mode, sub-mode, and speaker predilection on style. *Communication Monographs, 41*(2), 85–118.
- Bowey, J. A. (1995). Socioeconomic status differences in preschool phonological sensitivity and first-grade reading achievement. *Journal of Educational Psychology, 87*(3), 476–487. doi:10.1037/0022-0663.87.3.476
- Chafe, W., & Tannen, D. (1987). The relation between written and spoken language. *Annual Review of Anthropology, 16*(1), 383–407. doi:10.1146/annurev.an.16.100187.002123
- Chall, J. S., & Jacobs, V. A. (1983). Writing and reading in the elementary grades: Developmental trends among low SES children. *Language Arts, 60*(5), 617–626.
- Chall, J. S., & Jacobs, V. A. (2003). The classic study on poor children's fourth-grade slump. *American Educator, 27*(1), 14–15.
- Charalabidis, Y. N., Loukis, E., Androutopoulou, A., Karkaletsis, V., & Triantafillou, A. (2014). Passive crowdsourcing in government using social media. *Transforming Government: People, Process and Policy, 8*(2), 283–308.
- Collins, S. E. (2016). Associations Between Socioeconomic Factors and Alcohol Outcomes. *Alcohol Research: Current Reviews, 38*(1), 83–94. PMID:27159815
- Cronn-Mills, D., & Croucher, S. M. (2001). *Judging the Judges: An Analysis of Ballots in Impromptu and Extemporaneous Speaking*. Academic Press.
- D'angiulli, A., Siegel, L. S., & Maggi, S. (2004). Literacy instruction, SES, and word-reading achievement in English-language learners and children with English as a first language: A longitudinal study. *Learning Disabilities Research & Practice, 19*(4), 202–213. doi:10.1111/j.1540-5826.2004.00106.x
- DeVito, J. A. (1965). Comprehension factors in oral and written discourse of skilled communicators. *Communication Monographs, 32*(2), 124–128.

- Dohrenwend, B. P. (1990). Socioeconomic status (SES) and psychiatric disorders. *Social Psychiatry and Psychiatric Epidemiology*, 25(1), 41–47. PMID:2406949
- Doan, A., Ramakrishnan, R., & Halevy, A. Y. (2011). Crowdsourcing systems on the world-wide web. *Communications of the ACM*, 54(4), 86–96. doi:10.1145/1924421.1924442
- Emig, J. (1977). Writing as a mode of learning. *College Composition and Communication*, 28(2), 122–128. doi:10.2307/356095
- Fernald, A., Marchman, V. A., & Weisleder, A. (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental Science*, 16(2), 234–248. doi:10.1111/desc.12019 PMID:23432833
- Flaounas, I., Ali, O., Lansdell-Welfare, T., De Bie, T., Mosdell, N., & Lewis, J. (2013). Research methods in the age of digital journalism: Massive-scale automated analysis of news-content—topics, style, and gender. *Digital Journalism*, 1(1), 102–116. doi:10.1080/21670811.2012.714928
- Flesch, R. (1948). A new readability yardstick. *The Journal of Applied Psychology*, 32(3), 221–233. doi:10.1037/h0057532 PMID:18867058
- Goldberg, S., Fruchter, E., Davidson, M., Reichenberg, A., Yoffe, R., & Weiser, M. (2011). The relationship between the risk of hospitalization for schizophrenia, SES, and cognitive functioning. *Schizophrenia Bulletin*, 37(4), 664–670. doi:10.1093/schbul/sbr047 PMID:21602306
- Hanscombe, K. B., Trzaskowski, M., Haworth, C. M., Davis, O. S., Dale, P. S., & Plomin, R. (2012). Socioeconomic status (SES) and children’s intelligence (IQ): In a UK-representative sample SES moderates the environmental, not genetic, effect on IQ. *PLoS One*, 7(2), e30320. doi:10.1371/journal.pone.0030320 PMID:22312423
- Heppt, B., Haag, N., Böhme, K., & Stanat, P. (2015). The role of academic-language features for reading comprehension of language-minority students and students from low-SES families. *Reading Research Quarterly*, 50(1), 61–82. doi:10.1002/rrq.83
- Irani, L. (2017). Amazon Mechanical Turk. *The Blackwell Encyclopedia of Sociology*, 1–3. doi:10.1002/9781405165518.wbeos0994
- Kawachi, I., Kennedy, B. P., & Glass, R. (1999). Social capital and self-rated health: A contextual analysis. *American Journal of Public Health*, 89(8), 1187–1193. doi:10.2105/AJPH.89.8.1187 PMID:10432904
- Kellogg, R. T., & Raulerson, B. A. (2007). Improving the writing skills of college students. *Psychonomic Bulletin & Review*, 14(2), 237–242. doi:10.3758/BF03194058 PMID:17694907
- Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel*. Academic Press.
- Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel*. Research Branch Report 8–75. Chief of Naval Technical Training: Naval Air Station Memphis.
- Mailloux, S. L., Johnson, M. E., Fisher, D. G., & Pettibone, T. J. (1995). How reliable is computerized assessment of readability? *Computers in Nursing*, 13, 221–221. PMID:7585304
- Mayes, S. D., & Calhoun, S. L. (2011). Impact of IQ, age, SES, gender, and race on autistic symptoms. *Research in Autism Spectrum Disorders*, 5(2), 749–757. doi:10.1016/j.rasd.2010.09.002
- Nippold, M. A., Frantz-Kaspar, M. W., Cramond, P. M., Kirk, C., Hayward-Mayhew, C., & MacKinnon, M. (2014). Conversational and narrative speaking in adolescents: Examining the use of complex syntax. *Journal of Speech, Language, and Hearing Research: JSLHR*, 57(3), 876–886. doi:10.1044/1092-4388(2013)13-0097 PMID:24167229
- Odell, L. (1980). The process of writing and the process of learning. *College Composition and Communication*, 31(1), 42–50. doi:10.2307/356632
- Paniagua, J., & Korzynski, P. (2017). Social Media Crowdsourcing. *Encyclopedia of Creativity, Invention, Innovation and Entrepreneurship*, 1-4. doi:10.1007/978-1-4614-6616-1_200009-1

- Parameswaran, M., & Whinston, A. B. (2007). Social computing: An overview. *Communications of the Association for Information Systems*, 19(1), 37.
- Pew Research Report. (n.d.). <https://www.pewresearch.org/fact-tank/2019/02/15/the-changing-face-of-congress/>
- Piko, B., & Fitzpatrick, K. M. (2001). Does class matter? SES and psychosocial health among Hungarian adolescents. *Social Science & Medicine*, 53(6), 817–830. doi:10.1016/S0277-9536(00)00379-8 PMID:11511056
- Rajput, A. (2020). Natural Language Processing, Sentiment Analysis, and Clinical Analytics. In *Innovation in Health Informatics* (pp. 79-97). Academic Press.
- Rajput, A., & Ahmed, S. (2018a). Big Data and Social/Medical Sciences: State of the Art and Future Trends. *IACHS 2018*. Available at arXiv preprint arXiv:1902.00705.
- Rajput, A., & Ahmed, S. (2018b). Making a case for Social Media Corpus to detect Depression. *IACHSS 2018*. Available at arXiv preprint arXiv:1902.00702.
- Rajput, A. E. (2019, April). Using Crowdsourcing to Identify a Proxy of Socio-economic Status. In *The International Research & Innovation Forum* (pp. 479–486). Springer. doi:10.1007/978-3-030-30809-4_44
- Rajput, A., & Rotenstreich, S. (2004). Making A Case for Resource Management in a P2P Environment. In *IKE* (pp. 475-484). Academic Press.
- Schvaneveldt, R. W., Meyer, D. E., & Becker, C. A. (1976). Lexical ambiguity, semantic context, and visual word recognition. *Journal of Experimental Psychology. Human Perception and Performance*, 2(2), 243–256. doi:10.1037/0096-1523.2.2.243 PMID:1271030
- Si, L., & Callan, J. (2001, October). A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management* (pp. 574-576). ACM. doi:10.1145/502585.502695
- Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *The Journal of Documentation*, 28(1), 11–21. doi:10.1108/eb026526
- Stockmeyer, N. O. (2009). Using Microsoft Word’s readability program. *Michigan Bar Journal*, 88, 46.
- Veenstra, G. (2000). Social capital, SES and health: An individual-level analysis. *Social Science & Medicine*, 50(5), 619–629. doi:10.1016/S0277-9536(99)00307-X PMID:10658843
- Vitaro, F., Arseneault, L., & Tremblay, R. E. (1999). Impulsivity predicts problem gambling in low SES adolescent males. *Addiction (Abingdon, England)*, 94(4), 565–575. doi:10.1046/j.1360-0443.1999.94456511.x PMID:10605852
- Wang, L. W., Miller, M. J., Schmitt, M. R., & Wen, F. K. (2013). Assessing readability formula differences with written health information materials: Application, results, and recommendations. *Research in Social & Administrative Pharmacy*, 9(5), 503–516. doi:10.1016/j.sapharm.2012.05.009 PMID:22835706
- Wazny, K. (2017). “Crowdsourcing” ten years in A review. *Journal of Global Health*, 7(2), 020602. doi:10.7189/jogh.07.020601 PMID:29302322
- Weinberg, B. D., & Williams, C. B. (2006). The 2004 US Presidential campaign: Impact of hybrid offline and online ‘meetup communities. *Journal of Direct, Data and Digital Marketing Practice*, 8(1), 46–57. doi:10.1057/palgrave.ddmp.4340552
- Youyou, W., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences of the United States of America*, 112(4), 1036–1040. doi:10.1073/pnas.1418680112 PMID:25583507
- Tahir, M., Usman, M., Muhammad, F., Khan, I., Idrees, M., Irfan, M., & Glowacz, A. (2020). Evaluation of Quality and Readability of Online Health Information on High Blood Pressure Using DISCERN and Flesch-Kincaid Tools. *Applied Sciences (Basel, Switzerland)*, 10(9), 3214. doi:10.3390/app10093214
- Baier, S. L. (2021). Readability, Complexity, Flesch-Kincaid, Policy Analytics, Law, Statutes, Plain Language. Academic Press.

Somasundaram, M., Novak, C. B., Zuker, R. M., & Borschel, G. H. (2021). Facial Paralysis Online Educational Resources: Readability and Benefit to Patient Education. *Plastic and Reconstructive Surgery*, *148*(2), 10–1097. doi:10.1097/PRS.00000000000008164 PMID:34254962

Zhang, D., Earp, B. E., Kilgallen, E. E., & Blazar, P. (2021). Readability of Online Hand Surgery Patient Educational Materials: Evaluating the Trend Since 2008. *The Journal of Hand Surgery*.

ENDNOTES

- ¹ <https://www.webfx.com/tools/read-able/flesch-kincaid.html>
- ² <https://www.reddit.com>
- ³ <https://www.merriam-webster.com/dictionary/crowdsourcing>
- ⁴ <https://www.financialsamurai.com/the-110th-rule-for-car-buying-everyone-must-follow/>
- ⁵ One Senator did not have a Twitter handle or we were unable to locate it
- ⁶ <https://www.carmax.com/articles/which-car-brands-have-oldest-youngest-buyers>

Samara Ahmed finished her Psychiatry Residency from Tufts University and her fellowship in Child Psychiatry from NYU. Currently she is a faculty at College of Medicine at King Abdulaziz University in Jeddah, Saudi Arabia. Her research interests span application of Machine Learning algorithms in the field of psychiatry.

Adil Rajput is an Assistant Professor in the Department of Computer Science at Effat University. Prior to joining Academia, he has served in various capacities in the industry focusing on delivering Enterprise IT and cyber security solutions.

Akila Sarirete is Dean of Effat College of Engineering Dean at Effat University. She received her PhD degree in Computer Science and Knowledge Management. Her main research interests are in artificial intelligence, knowledge management, communities of practice, machine learning, big data, and service-oriented architectures. She presented her research work in several conferences in different countries. Dr. Sarirete has a vast experience in software development industry and software engineering. She is interested in engineering education, innovation, smart cities and villages especially, the human aspect and the collaborative work.