Profile Analysis for Cryptocurrency in Social Media

Husnu S. Narman[*] narman@marshall.edu

Alymbek Damir Uulu[*] damiruulu@marshall.edu

Jinwei Liu jliu@ist.ucf.edu

[*]Weisberg Division of Computer Science, Marshall University, Huntington, WV 25755
Institute for Simulation and Training, University of Central Florida, Orlando, FL 32826

***Abstract*—Blockchain ushers in a new era for the global financial system with the advent of digital currency (cryptocurrency), and its impact can be felt in many related industries. Because of its possible applications, cryptocurrency draws significant attention from researchers. Although there are a number of risks (e.g., speculation, 51% attack) related to cryptocurrency, billions of dollars are invested in them, because of transparency, traceability, low transaction cost, and highly profitable potential. In December 2017, the most famous cryptocurrency, Bitcoin, has reached almost $20,000.00 per coin. Such short-term, high gain potential attracts many new small investors. However, speculative movements raise many questions related to safety and privacy, just to name a few. In order to understand public opinion about cryptocurrency and to protect small investors financial interests, sentiment analysis can be done by using social media activities of individuals who are interested or investing in cryptocurrencies. One of the most important steps in the analysis is to understand the profiles of the users. Therefore, in this paper, we determine education levels of investors or users who are interested in eight cryptocurrencies by using seven readability techniques on Reddit comments as a part of profiling. Results show that the education levels of users are approximately 60% in middle school, 30% in high school, and 10% in other levels according to the average of the seven readability technique results. The results and analysis, which are provided in this paper, help new investors and developers to obtain profile information about the users who are interested or investing in cryptocurrency.**

***Index Terms*—Cryptocurrency; Social Media; Profile Analysis; Readability.**

I. Introduction

Blockchain ushers in a new era for the global financial system with the advent of digital currency, and its impact can be felt in other related industries [1]. With its ongoing developments and increased applications of the blockchain, it draws significant attention from researchers. One of the most important benefits of the blockchain, particularly with financial systems is the use of cryptocurrency. Although there are a number of risks (e.g., taxation, speculation, pseudo-anonymity, and 51% attack) related to cryptocurrency [2], billions of dollars are invested in it [3], [4] largely due to its permanent transparency, traceability, low transaction cost, pseudo-anonymity transactions [1], and high profitable profiles. In December 2017, the most famous cryptocurrency, Bitcoin has reached almost $20,000.00 per coin [4] which is ten times higher comparing to the previous year. Such a short-term, high gain venture attracts many new small investors. However, speculative movements cause cryptocurrency bans [5] and bring a number of investment questions, such as safety and privacy, just to name a few. In order to understand the public opinion of the cryptocurrencies and also protect the new investors, sentiment analysis can be done by using social media activities of cryptocurrency related forums because small investors and interested parties follow and actively enroll in social media, including Twitter [6], Reddit [7], YouTube [8], and many other social media to get more information about the features of coins and future gain possibilities [9].

Text analysis in social media is widely used to understand the trends of users [10], [11]. As a result of text analysis of users' comments on cryptocurrency subjects, researchers have identified a strong interaction between the social media sentiment and the Bitcoin price, and a tendency for investors to overreact to the news on social media within a short period [9]. However, a social marketing strategy can negatively affect the investors [12] in the long term. Therefore, it is important to analyze the and identify profiles of cryptocurrency investors and who are interested in it in order to protect new investors from the financial losses and provide profile information to the new investors and interested parties.

There are several works which analyze cryptocurrency in terms of security, privacy, applications, usability, regulations, and technology [1]–[3], [13], [14]. Although there is no text mining analysis specifically on profile analysis of cryptocurrency activities, there are text mining research works on price predictions [9], [15], [16]. In [15], social network data is analyzed in order to better understand the factors underlying the price and other trends in emerging cryptocurrency markets. Similarly, in [9], social network data is studied to understand the relation between bitcoin price and social activities. In [16], keywords are extracted from Bitcoin-related user comments posted on the online forum to analytically predict the price and extent of transaction fluctuation of the currency. The previous works mostly focus on price predictions of cryptocurrencies, especially Bitcoin by using text analysis. On the other hand, this paper *aims* to provide profile information. As an initial component of the profile information, we are interested in education levels of the users who are active in social media which related to cryptocurrencies.

The *objective* of this paper is to analyze the education levels of the users who are active in eight cryptocurrency subreddits (Bitcoin, Bitcoin Cash, Dash, Ether, Litecoin, Lumen, Monero, and Ripple) by using users' comments for each coin subreddit on Reddit. The key *contributions* of this paper can be listed as follows:

- The eight cryptocurrencies are investigated in terms of user education levels.
- The seven text readability techniques are used to analyze more than 50,000 users' comments for the aforementioned cryptocurrencies.

The *results* show that the education levels of users are approximately 60% in middle school, 30% in high school, and 10% in other levels according to the average of the seven readability technique results while the education levels of the same groups are approximately 50% in high school, 35% in college, and 15% in other levels according to Fog Scale readability technique. The results and analysis, which are provided in this paper, help new investors and developers to obtain profile information about the users who are interested or investing in cryptocurrency.

The rest of the paper is organized as follows: In Section II, the system model and assumptions are explained. In Section IV, the text readability techniques are discussed. In Section V, results are presented, and finally, Section VI has the concluding remarks with future works.

II. System Model

In this section, we explain the data gathering model.

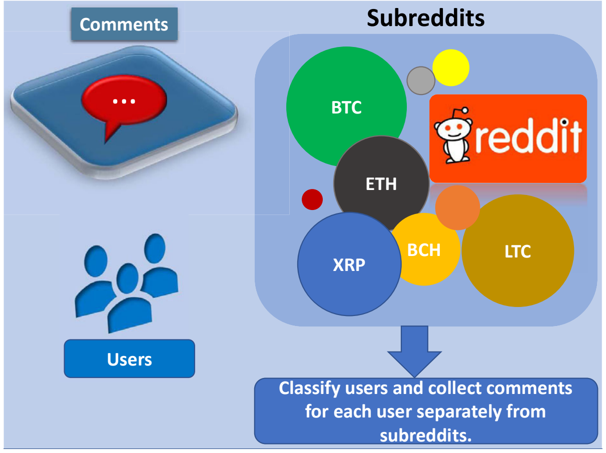

Comments
Subreddits
BTC
reddit
ETH
XRP
BCH
LTC
Users
Classify users and collect comments for each user separately from subreddits.

Fig. 1: Data gathering model from Reddit for eight cryptocurrencies.

Fig. 1 shows the data collection process for each coin from Reddit. Reddit can have one or more subreddits for each cryptocurrency, and each cryptocurrency can have a number of posts in each subreddit. Moreover, each post can have many comments from a number of users because users tend to respond to posts that match their interests. We use ten to seventy top posts for each cryptocurrency to collect distinct usernames. Then, comments of each user are collected according to usernames. We assume that if a user comments on a subreddit post which belongs to a coin, either user is an investor or is interested in the coin. It is possible that the user is interested in the coin, but has not invested in it. Moreover, because of the informal structure of the comments (no or missing punctuation, shortened words and so forth), the obtained results approximate to the education levels of users.

III. Readability Indices

In this section, we explain the used readability test techniques to test the collected comments to identify the education levels of the users.

A. Flesch–Kincaid Readability

Flesch formula has been developed [17] in order to measure the readability of written texts by giving scores from 0 to 100. Kincaid formula has also been developed by using the same concepts (syllabus, words, and sentences) but uses grade instead of scores to explain the readability of texts [18]. Table I shows the readability grades for Flesch–Kincaid.

TABLE I: Flesch-Kincaid readability with the score, grades, and descriptions.

Score	Grades	Description
$\geq$ 90	5th	Very easy
90-80	6th	Easy
80-70	7th	Fairly easy
70-60	8th-9th	Standard
60-50	10th-12th	Fairly difficult
50-30	College	Difficult
$\leq$ 30	Graduated	Very difficult

B. Dale-Chall Readability

Dale-Chall readability uses words, difficult words, and sentences to provide a score to test the readability of text [19]. Dale-Chall identifies the easiness of words with a list of 3,000 words that are expected to be known by fourth-grade in the USA. The words which are not on the list are accepted as difficult. Table II shows the readability scores with grades for Dale-Chall.

TABLE II: Dale-Chall readability scores with grades.

Index	Grades
$\leq$ 4.9	Understood by 4th-grades or lower
5.0-5.9	Understood by 5th or 6th-grades
6.0-6.9	Understood by 7th or 8th-grades
7.0-7.9	Understood by 9th or 10th-grades
8.0-8.9	Understood by 11th or 12th-grades
9.0-9.9	Understood by 13th to 15th-grades (college)

C. The Fog Scale (Gunning Fog)

Fog Scale has been developed to approximately predict the education level of a person to understand a given text on the first reading [20] and is usually used to verify whether text can be easily understood by readers. It uses words, complex words, and sentences to measure the readability of text. Complex words are the words which consist of three or more syllables that do not include the common suffix, composed of two or more words, and proper nouns. Fog scale scores and grades are shown in Table III.

TABLE III: Fog Scale, SMOG and Coleman-Liau indices with grades.

Index	Grade
5	5th-grade
6	6th-grade
7	7th-grade
8	8th grade
9-12	High schools (9 freshmen,..., 12 senior)
13-16	College (13 freshmen,..., 16 senior)
$\geq$ 17	Graduated

D. Automated Readability Index

Automated Readability Index is developed to understand the required grade level to comprehend written texts. It uses characters, words, and sentences while measuring texts. Table IV shows Automated Readability Index with school levels.

TABLE IV: Automated Readability index for education levels in terms of schools.

Grades	Possible schools
1	Kindergarten students
2-4	Elementary School students
5-8	Middle School students
9-12	High School students
13	College students
$\geq$ 14	Graduated

E. Simple Measure of Gobbledygook (SMOG)

Simple Measure of Gobbledygook (SMOG) is developed to find the required education level to understand written texts [21], and is commonly used in health messages. SMOG uses sentences and polysyllables to test texts. Table III also shows the SMOG Index for grade levels.

F. Coleman-Liau Index

Coleman-Liau Index is created to test the readability of the text by using the average number of characters and sentences per one hundred words [22] instead of using syllables. Table III also shows the Coleman-Liau Index for grade levels.

G. Linsear Write

The Linsear Write formula is developed by the United States Air Force to measure the readability of their technical manuals. It is based on sentences and words that have three or more syllables. The words which are formed by one or two syllables are accepted as easy, and the words which are formed by three or more syllables are accepted as hard words. Table III is used to compare Linsear with other readability techniques.

H. Difficulty of Words

The difficulty of words is also used as a measurement in this paper to show what percentage of users, which use difficult words in their comments. The list has 3,000 easy words, which are used for Dale-Chall Readability techniques. The score of the difficulty of words is changing between 0 and 100 (0 presents there are no difficult words, 100 means very complex)

I. Average of Seven Readability Techniques

Each readability techniques which are explained above have similar or distinct result while showing the education levels of users. Therefore, the average of all results from the above readability techniques (except the difficulty of words) are measured to provide a common result.

IV. Selected Cryptocurrencies

In this section, we explain the selected cryptocurrencies to test in this paper. Although there are more than 1000 coins and tokens, we have selected Bitcoin (BTC), Bitcoin Cash (BCH), Litecoin (LTC), Ether (ETH), Lumen (XLM), Monero (XMR), Dash (DASH), and Ripple (XRP) to analyze.

A. Bitcoin (BTC)

Bitcoin was the first proposed decentralized cryptocurrency to eliminate the financial institutions while making transactions or payments directly from one party to another by using peer-to-peer network [23]. It is a milestone for cryptocurrency and widely used by the cryptocurrency community.

B. Bitcoin Cash (BCH)

Bitcoin Cash is forked from Bitcoin to increase the transaction speed of Bitcoin by reducing transaction fees in 2017 [24]. Although there are a number of the forked coin from Bitcoin, Bitcoin Cash is widely used by the cryptocurrency community, and reached one-third of Bitcoin volume in a day.

C. Litecoin (LTC)

Litecoin is also a decentralized currency. The main difference between Bitcoin and Litecoin is the mining algorithms. While Bitcoin is SHA-256 based, Litecoin is Scrypt based cryptocurrency which authenticates blocks of transaction data [25]. SHA-2 series is developed by the United States National Security Agency. Scrypt is developed for use in the Tarsnap online backup system [26].

D. Ether (ETH)

Contrast to Bitcoin, Ethereum is a platform and is not just to support digital currency but supports the creation of applications [27]. It can also be called Blockchain 2.0 because of expanding traditional blockchain boundaries with smart contracts and crowdsourcing. The applications, which created on this platform, can use Ether token to run and trade [27].

E. Ripple (XRP) and Lumen (XLM)

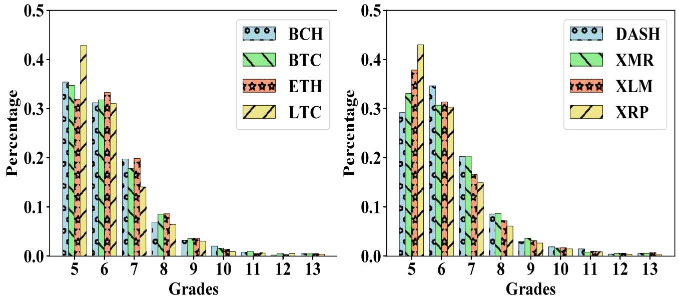
Ripple is a payment settling and currency exchange which developed for banks and payment networks. It aims to provide a system for direct transfer of assets (e.g., money.) in real-time (3 seconds) with the lower price. It is an alternative for the SWIFT payment system. Ripple uses a distributed consensus ledger which using a network of validating servers instead of the blockchain, and its token is XRP [28]. Stellar is similar to Ripple but uses Lumen as a token. Although Banks and other organizations can also use Stellar, its initial target was citizens and small organizations to be able to transfer assets in real-time with the cheaper price [29].

F. Monero (XMR) and Dash (DASH)

Monero and Dash were developed to address the traceable concerns that Bitcoin has. With Monero and Dash, the transactions which have been made hard to link any real identity [30], [31].

V. Results

In this section, we present the results which we obtained by using seven readability techniques.



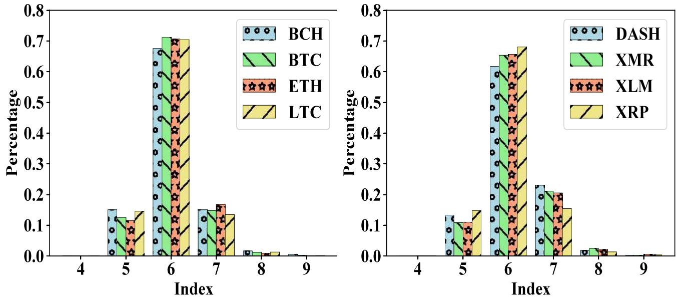
(a) BCH, BTC, ETH, and LTC.

(b) DASH, XMR, XLM, and XRP.

Fig. 2: The percentage of users' grades for eight cryptocurrencies according to Flesch-Kincaid readability.

A. Flesch–Kincaid Readability

Figures 2a and 2b show the Flesch-Kincaid Readability results for the selected cryptocurrencies. Although there are small differences between cryptocurrencies, the grades of users are mostly grouped under 5th, 6th and 7th grades, and approximately 80% of users are in 5th to 7th grades. The percentage of college students and the users graduated from a college are significantly lower.



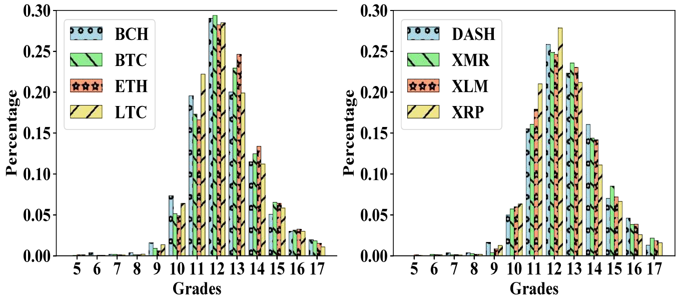
(a) BCH, BTC, ETH, and LTC.

(b) DASH, XMR, XLM, and XRP.

Fig. 3: The percentage of users' grades for eight cryptocurrencies according to Dale-Chall Readability.

B. Dale-Chall Readability

Figures 3a and 3b illustrate the Dale-Chall Readability results for the selected cryptocurrencies. Although there are slight differences between cryptocurrencies, the grades of users are grouped under 7th and 8th grades. According to the results, nearly 70% of users' grades are 7th and 8th. The percentage of college students and the users graduated from a college are almost 2%.



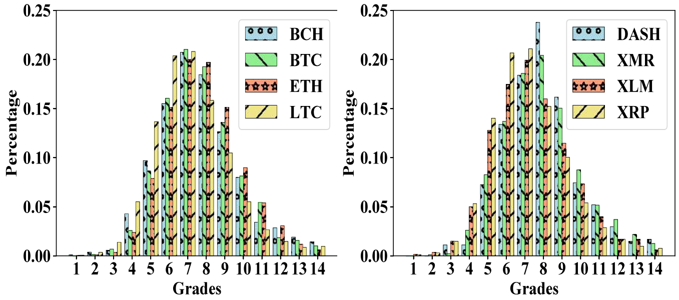
(a) BCH, BTC, ETH, and LTC.

(b) DASH, XMR, XLM, and XRP.

Fig. 4: The percentage of users' grades for eight cryptocurrencies according to Fog Scale (Gunning Fog).

C. The Fog Scale (Gunning Fog)

Figures 4a and 4b illustrate the Fog Scale results for the selected cryptocurrencies. The grades of 50% users are grouped under 9th to 12th grades. According to the results, almost 35% of users are college students.



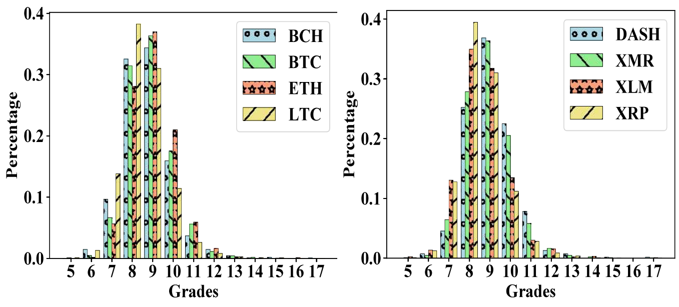
(a) BCH, BTC, ETH, and LTC.

(b) DASH, XMR, XLM, and XRP.

Fig. 5: The percentage of users' grades for eight cryptocurrencies according to Automated Readability Index.

D. Automated Readability Index

Figures 5a and 5b show the Automated Readability Index results for the selected cryptocurrencies. Although DASH shows different patterns, compared to other cryptocurrencies, it is observed that up to 65% of users are from 5 to 8 grades. Up to 30% of users are from 9 to 12 grades. Almost 2.5% of users are either in a college or graduated from a college.



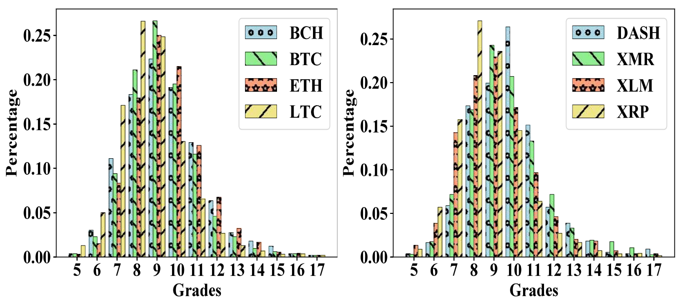
(a) BCH, BTC, ETH, and LTC.

(b) DASH, XMR, XLM, and XRP.

Fig. 6: The percentage of users' grades for eight cryptocurrencies according to SMOG.

E. Simple Measure of Gobbledygook (SMOG)

Figures 6a and 6b show the Simple Measure of Gobbledygook (SMOG) results for the selected cryptocurrencies. At least, 65% of users are in 8th and 9th grades. The percentage of users in a college or graduated from a college is significantly lower.



(a) BCH, BTC, ETH, and LTC.

(b) DASH, XMR, XLM, and XRP.

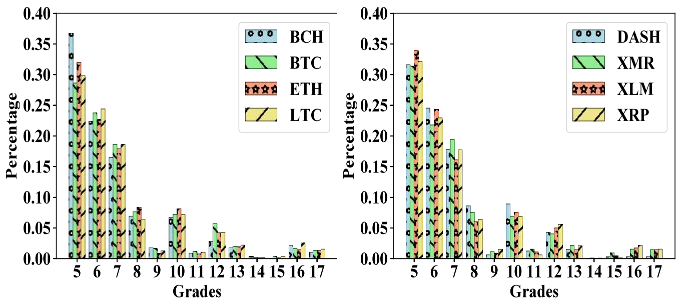
Fig. 7: The percentage of users' grades for eight cryptocurrencies according to Coleman-Liau Index.

F. Coleman-Liau Index

Figures 7a and 7b show the Coleman-Liau Index results for the selected cryptocurrencies. The result patterns are significantly different from SMOG results. Almost, 45% of users are in 8th and 9th grades, and the percentage of 10th-grade users are nearly 20% for BTC, BCH, and ETH while XRP and LTC users are almost 15% in 10th grade. DASH has the highest percentage for grade 10th with 25%. The percentage of users in a college or graduated from a college is significantly lower as similar to SMOG results.

G. Linsear Write

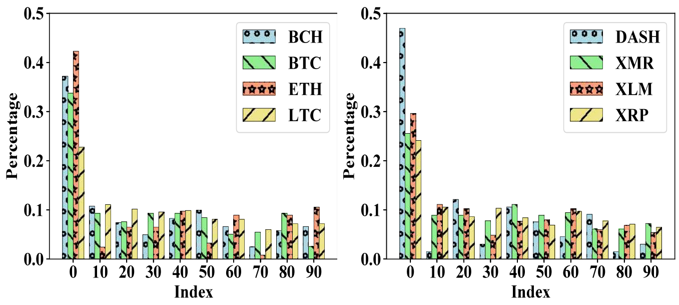
Figures 8a and 8b show the Linsear Write results for the selected cryptocurrencies. Although there are small differences between coins and tokens, the users are mostly grouped under 5th, 6th and 7th grades, and roughly 75% of users are in 5th, 6th and 7th grades. The percentage of college students and the users graduated from a college are significantly lower and almost 8%.



(a) BCH, BTC, ETH, and LTC.

(b) DASH, XMR, XLM, and XRP.

Fig. 8: The percentage of users' grades for eight cryptocurrencies according to Linsear Write.



(a) BCH, BTC, ETH, and LTC.

(b) DASH, XMR, XLM, and XRP.

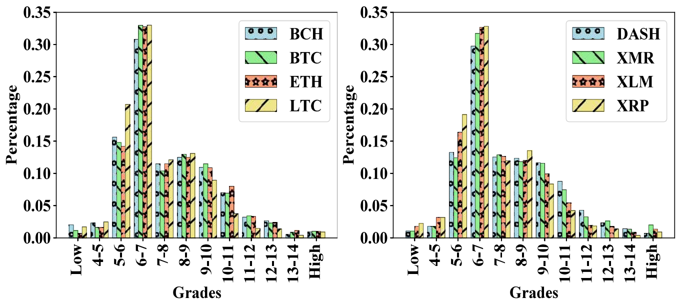
Fig. 9: The difficult word usage percentage of users for eight cryptocurrencies.

H. Difficulty of Words

Figures 9a and 9b show the difficulty word usage results for the selected cryptocurrencies. DASH posts have the lowest difficulty, and almost 50% of comments do not contain difficult words. On the other hand, roughly 40% of comments for BCH, BTC, and ETH do not contain difficult words. Approximately, 25% of XMR, XRP, XLM, and LTC comments are not difficult.

I. Standard Grades

Figures 10a and 10b show the grades of users by taking the average of all readability techniques except the difficulty of words. The results show that the education levels of 60% users are approximately in middle school (5, 6, 7, and 8). While at least 15% of users are in 5-6 grades, at least 35% of users are in 6-7 grades. Moreover, nearly 30% of users in high school (9, 10, 11, and 12), and roughly 10% of users are on other levels. It is important to note that LTC, XRP, and XLM results are slightly different from other cryptocurrencies.



(a) BCH, BTC, ETH, and LTC.

(b) DASH, XMR, XLM, and XRP.

Fig. 10: The percentage of users' grades for eight cryptocurrencies according to standard grade calculation (average).

J. Summary of Results

The results which have been obtained by using seven readability techniques on Reddit comments can be summarized as follows: (i) There are differences between the obtained results from seven readability techniques. The most significant difference is the Fog Scale results which show that 50% of users are grouped under 9th to 12th grades, and almost 35% of users are college students. (ii) The average results of seven readability techniques show that the education levels of users are approximately 60% in middle school, nearly 30% in high school, and 10% in other levels. (iii) The results which are obtained for eight cryptocurrencies are distinct.

VI. Conclusion and Future Works

In this paper, we analyze the education levels of users who are interested in cryptocurrencies in order to provide information about user profiles to new investors. To obtain the education level information, we use Reddit.com, collect and classify the gathered comments data from subreddits of eight cryptocurrencies, and analyze the data according to seven text readability techniques. We find out that the results which we obtained from distinct readability techniques have differences. Therefore, the average results of seven readability techniques are measured to provide a common result. According to the average of the results of seven readability techniques, the education levels of users are approximately 60% in middle school, 30% in high school, and 10% in other levels. However, the education levels of users are approximately 50% in high school, 35% in college, and 15% in other levels according to Fog Scale readability technique. The results and analysis, which are provided in this paper help new investors and developers to have profile information about users who are interested in cryptocurrencies.

In the future, we would like to extend this work by obtaining not only education levels but also behavioral characteristics of users from multiple social media platforms.

References

[1] F. Tschorsch and B. Scheuermann, "Bitcoin and beyond: A technical survey on decentralized digital currencies," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 2084–2123, thirdquarter 2016.

[2] C. G. Harris, "The risks and dangers of relying on blockchain technology in underdeveloped countries," in *NOMS 2018 - 2018 IEEE/IFIP Network Operations and Management Symposium*, April 2018, pp. 1–4.

[3] T. Salman, M. Zolanvari, A. Erbad, R. Jain, and M. Samaka, "Security services using blockchains: A state of the art survey," *IEEE Communications Surveys Tutorials*, 2018.

[4] Bitcoin. Accessed: Aug. 14, 2018. [Online]. Available: https://coinmarketcap.com/currencies/bitcoin/

[5] J. Russell. China has banned ICOs. Accessed: Sep. 4, 2017. [Online]. Available: https://techcrunch.com/2017/09/04/chinas-central-bank-has-banned-icos/

[6] Twitter. [Online]. Available: https://www.twitter.com/

[7] Reddit. [Online]. Available: https://www.reddit.com/

[8] Youtube. [Online]. Available: https://www.youtube.com/

[9] V. Karalevicius, N. Degrande, and J. D. Weerdt, "Using sentiment analysis to predict interday bitcoin price movements," *The Journal of Risk Finance*, vol. 19, no. 1, pp. 56–75, Sep. 2018.

[10] R. Irfan, C. K. King, D. Grages, S. Ewen, S. U. Khan, S. A. Madani, J. Kolodziej, L. Wang, D. Chen, A. Rayes *et al.*, "A survey on text mining in social networks," *The Knowledge Engineering Review*, vol. 30, no. 2, pp. 157–170, 2015.

[11] S. Zhou, H. Jeong, and P. A. Green, "How consistent are the best-known readability equations in estimating the readability of design standards?" *IEEE Transactions on Professional Communication*, vol. 60, no. 1, pp. 97–111, 2017.

[12] M. Corstjens and A. Umblijs, "The power of evil: The damage of negative social media strongly outweigh positive contributions," *Journal of Advertising Research*, vol. 52, no. 4, pp. 433–449, 2012.

[13] A. Goranovic, M. Meisel, L. Fotiadis, S. Wilker, A. Treytl, and T. Sauter, "Blockchain applications in microgrids an overview of current projects and concepts," in *43rd Annual Conference of the IEEE Industrial Electronics Society*, Oct 2017, pp. 6153–6158.

[14] M. C. K. Khalilov and A. Levi, "A survey on anonymity and privacy in bitcoin-like digital cash systems," *IEEE Communications Surveys Tutorials*, 2018.

[15] M. Laskowski and H. M. Kim, "Rapid prototyping of a text mining application for cryptocurrency market intelligence," in *IEEE 17th International Conference on Information Reuse and Integration (IRI)*, July 2016, pp. 448–453.

[16] Y. B. Kim, J. Lee, N. Park, J. Choo, J.-H. Kim, and C. H. Kim, "When bitcoin encounters information in an online forum: Using text mining to analyse user opinions and predict value fluctuation," *PloS one*, vol. 12, no. 5, p. e0177630, 2017.

[17] R. F. Flesch, *How to Write Plain English*.

[18] G. M. McClure, "Readability formulas: Useful or useless?" *IEEE Transactions on Professional Communication*, vol. PC-30, no. 1, pp. 12–15, March 1987.

[19] E. Dale and J. S. Chall, "A formula for predicting readability," *Educational Research Bulletin*, vol. 27, 1948.

[20] R. Gunning, *The technique of clear writing*, 1952.

[21] G. H. Mc Laughlin, "SMOG grading-a new readability formula," *Journal of reading*, vol. 12, no. 8, pp. 639–646, 1969.

[22] M. Coleman and T. L. Liau, "A computer readability formula designed for machine scoring." *Journal of Applied Psychology*, vol. 60, no. 2, p. 283, 1975.

[23] S. Nakamoto, "Bitcoin: A peer-to-peer electronic cash system."

[24] Bitcoin cash. [Online]. Available: https://www.bitcoincash.org/

[25] Litecon. [Online]. Available: https://litecoin.com/

[26] Scrypt. [Online]. Available: http://www.tarsnap.com/scrypt.html

[27] Ethereum. [Online]. Available: https://www.ethereum.org/

[28] Ripple, xrp. [Online]. Available: https://ripple.com/

[29] Stellar lumens. [Online]. Available: https://www.stellar.org/

[30] Monero. [Online]. Available: https://getmonero.org/

[31] Dash. [Online]. Available: https://www.dash.org/