Evaluating Posts on the Steemit Blockchain: Analysis on Topics Based on Textual Cues

Kristina Kapanova
Trinity College Dublin
Dublin, Ireland
kapanovk@tcd.ie

Barbara Guidi
Department of Computer Science, University of Pisa
Pisa, Italy
guidi@di.unipi.it

Andrea Michienzi
Department of Computer Science, University of Pisa
Pisa, Italy
andrea.michienzi@di.unipi.it

Kevin Koidl
Trinity College Dublin
Dublin, Ireland
kevin.koidl@scss.tcd.ie

ABSTRACT

Online Social Networking platforms (OSNs) are part of the people's everyday life answering the deep-rooted need for communication among humans. During recent years, a new generation of social media based on blockchain became very popular, bringing the power of the technology to the service of social networks. Steemit is one such and employs the blockchain to implement a rewarding mechanism, adding a new, economic, layer to the social media service. The reward mechanism grants virtual tokens to the users capable of engaging other users on the platform, which can be either vested in the platform for increased influence or exchanged for fiat currency. The introduction of an economic layer on a social networking platform can seriously influence how people socialize. In this work, we tackle the problem of understanding how this new business model conditions the way people create contents. We performed term frequency and topic modelling analyses over the written contents published on the platforms between 2017 and 2019. This analysis lets us understand the most common topics of the contents that appear in the platform. While personal mundane information still appears, along with contents related to arts, food, travels, and sport, we also see emerging a very strong presence of contents about blockchain, cryptocurrency and, more specifically, on Steemit itself and its users.

CCS CONCEPTS

• **Human-centered computing** → **Social network analysis**; **Social media**; • **Computing methodologies** → *Natural language processing*.

KEYWORDS

Steemit, blockhain social networks, NLP

ACM Reference Format:
Kristina Kapanova, Barbara Guidi, Andrea Michienzi, and Kevin Koidl. 2020. Evaluating Posts on the Steemit Blockchain: Analysis on Topics Based on Textual Cues. In *6th EAI International Conference on Smart Objects and Technologies for Social Good (GoodTechs '20), September 14–16, 2020, Antwerp, Belgium.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3411170.3411248

GoodTechs '20, September 14–16, 2020, Antwerp, Belgium

ACM ISBN 978-1-4503-7559-7/20/09.
https://doi.org/10.1145/3411170.3411248

1 INTRODUCTION

With the rise of Online Social Networks (OSNs) in the past decade, people can share information now, more than ever, on an unprecedented scale and reach. Users share on social networks not only important news but also mundane information from their daily lives, including daily meals, fitness regimes, thoughts and feelings, recommendations and others. Many of the social networks do not provide direct financial rewards, thus users share information to accumulate either collective social capital, or capitalize indirectly on their built-in audiences (i.e. influencers). Understandably, this sharing activity has been extensively studied [7, 19, 20, 22] focusing not only on the consequences but also on the underlying network properties amplifying such behaviours.

With the recent revelations of the Cambridge Analytica scandal [5], Decentralized Online Social Networks (DOSNs) become of high interest for social media users [10]. During the last ten years, several DOSNs have been proposed [4, 9, 14, 15], but they did not have a big impact on the daily life of people. Thanks to the blockchain technology and the idea of rewarding valuable content, a new generation of DOSNs have been proposed, principally based on blockchain [13]. There are a variety of online social networks, relying on blockchain technology and monetization to encourage their users to create and share their content. The most well-known in Steemit[1]. It combines blockchain with social networking and blogging along with a monetary system, allowing participants in the platform to receive micropayments for the content they have generated. This is based on the number of votes their posts accrue, as well as the stake of the users who cast the votes. The incentive is geared toward the production of more content which should have an interesting enough topic to generate more votes. Besides content creators, users who identify (vote on) popular content are also rewarded for upvoting it. The latter is known as curation reward. The curation reward is higher for contents with a high number of upvotes and is granted for the most part to the first users who upvoted that content.

[1] https://steemit.com/

The main question about Blockchain Social Media concerns the behaviour of users and how the rewarding strategy affects their social behaviour. The platforms are based on rewarding strategies and the principal motivation is that it could prevent fake news and censorship [13]. However, the real benefit is still unclear.

In this paper, we provide a text analysis of the social content of Steemit. We utilize recent advancements in machine-learning methods to extract knowledge from the published content on Steemit to understand how the reward mechanism affects the creation of contents. We focus only on the analysis of the written parts of the contents because it is the main part of the contents posted on Steemit. In detail, we put our effort into discovering what are the most recurring and important topics of the contents present in the platform. We also try to identify some topics relevant to the platform and define them by the means of the words that appear most frequently in the contents belonging to these topics. Our investigation yields several important results. Firstly, some well-defined topics of contents emerge from the analyses, which suggests us that the platform is used properly and its contents can be relevant for readers. Secondly, by studying the most representative words that define each topic we can understand what are the most recurring and important topics of contents that appear in the platform.

The rest of the paper is structured as follows. In Section 2, we provide a detailed description of the techniques used for our analysis, as well as a brief overview of the field. Section 3 contains information about the data collection method, pre-processing, and a general description of the dataset. Section 4 presents the obtained results and the underlying analysis, followed by concluding remarks and possible future directions.

2 METHODS AND STATISTICAL APPROACHES

Natural Language Processing (NLP) represents a collection of techniques to process human language (written and spoken) to extract valuable information for some purpose, whether language translation, extract knowledge from a variety of textual sources, to answer questions, carry out automatic conversations, predict hate speech, track sentiment and others. NLP considers not only text as a sequence of symbols but also relies on the hierarchical structure of language - words that form phrases, which are then encompassed in sentences [27] .

The sheer number of online content that is created daily has provided an indispensable unstructured research resource for scientists to explore a variety of social interactions. For instance, in [3] NLP techniques and social network analysis approaches were used to identify cultural networks in autism spectrum disorder Facebook groups. NLP has been applied in prediction whether certain Twitter messages will elicit responses [2]. Latent Dirichlet Allocation (LDA) has been used to establish latent communication connections in two scientific communities. Recurrent neural networks were used to establish adverse drug reactions based on Twitter posts [8]. Prediction of latent personal attributes and analysis of emotions of users through Twitter messages was demonstrated in [26]. NLP approaches have been used to infer people's mental states based on social media data [6]. Additionally, disaster response analysis based on social media posts has been performed in multiple studies based on both supervised and unsupervised machine learning approaches [1, 21, 24, 25]. In as much as our motivations stem from the interest of identifying the most common topics of the contents and the most recurring words for each topic, we employ several techniques in the present study, which are described below.

2.1 Term Frequency

Term frequency is a common approach to retrieve valuable information from a large textual corpus, often in unstructured form. The approach is used in community analysis [17], predicting the market [23], to understand emergency situational awareness from Twitter messages [28] among others. One should note, however, that term frequency, or the number of unique times a word has appeared in a text, is highly correlated to the text's length. Naturally, longer documents have higher term frequencies, than shorter ones. To mitigate this shortcoming, we can weight the term count where term counts will add up to one. Term frequency is combined with inverse document frequency analysis (collectively known as TF-IDF analysis) and is used often to understand what a document is about and how important words in a text are based on their frequency and weights scores.

2.2 Topic modelling and Analysis

LDA is one of the most popular and effective unsupervised topic modelling techniques [18] developed to identify latent topics from documents. Generally, LDA detects related topics from associated and co-occurring elements within a collection of documents. Each topic is defined by a distribution over words, where the ordering of words is ignored and that the words in each document are known. The distribution of topics altogether, as well as the distribution of topics for each document, is learned from the data. The topic distribution is then represented as a vector, which is used for computing the distance between the documents, which then provides information about their similarity.

LDA has proven successful in encapsulating knowledge from large corpora, as showed in [12], as well as to understand scientific directions (trends) in a field like biochemistry through the years [16]. A factor to consider when obtaining LDA-topics is what should be the number of topics to be extracted. LDA algorithms are unable to define on their own the number of topics in a given data. Therefore that number should be manually set beforehand and can depend on the situation since too small of topics can result on topics being too general and vague, while a big number will create more noisy topics [11].

3 DATA PREPARATION

For this work, we have extracted from the Steemit public blockchain API, published posts in the period $01 - January - 2017$ - $26 - September - 2019$. Once the data are collected, pre-processing and cleaning is performed on the data to remove incomplete, noisy or inconsistent entries from the set. We have taken specific steps to preprocess the data and make it suitable for our analysis. This includes the conversion of all collected posts to $UTF-8$ format, the removal of all HTML tags, converting links to entities, which represent the domain but not the individual pages, convert all recognizable emoticons, convert all pictures or gifs to entities, represented as

jpeg or gif, respectively. Furthermore, we have removed all special characters and stop words, as well as white spaces. All text was then converted to lowercase.

Since the posts that were retrieved contain both English and non-English posts, once the data were cleaned, the language for each post was established. Table 1 shows the total number of posts collected for each year, how many posts remain in the dataset once the pre-processing step has been finalized and the number of languages detected in the cleaned data. Posts occurred in a total of 43 languages. For the following analysis, we are going to focus only on English language posts.

Year	# Posts	# English Posts	Lang
2017	32990	28823	40
2018	32895	28593	36
2019	26654	21323	43

Table 1: Amount of posts collected for each year.

3.1 Users

The number of unique authors of posts in the collected data (2017 – 2019) is 9, 252. Once again, the user-specific data collected from the Steemit API was preprocessed and the cleaned data consists of 9, 245 entries. Since the number of posts per users are in extreme (the user with most posts has published 2, 303, 298, and there are users, who have posted only once), to account for these extremes, Figure 1 presents the histogram for posts per users in a log scale.

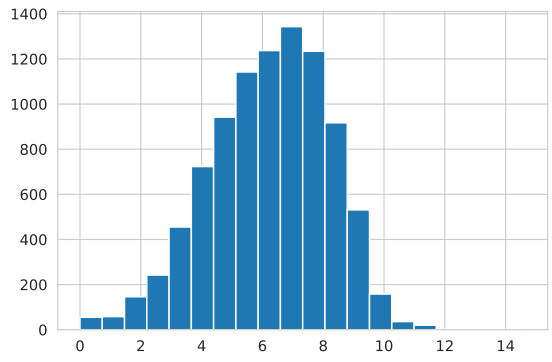

Figure 1: The data for number of posts is represented in log form in the histogram

3.2 Categories

Users can apply a single category, or tag, to each post they create as so to make easier for the contents to be categorized. The number of categories is 2, 509 for 2017, 2, 610 for 2018 and 2, 022 for 2019 respectively. Figure 2 shows the top 10 categories users published under for the entire period. It is noticeable that there is a mix of categories related to human activities, such as *photography*, *art*, *travel*, and *food*, together with some more related to the platform itself, such as *steemit*, *steem*, *bitcoin*. It is also worthwhile to notice that there are two categories probably related to the language

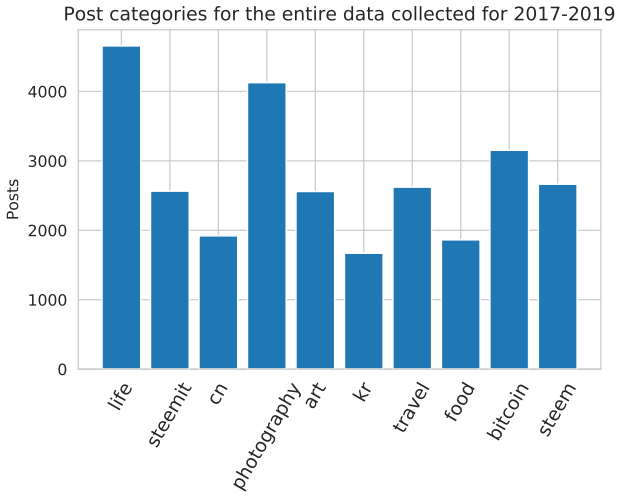

Figure 2: Top 10 **categories for all collected posts in the period** 2017 – 2019

used in the content: *kr* (Korean language code), and *cn* (Chinese language code). This behaviour may be linked to the fact that the two languages are not as present as English, so assigning them a category makes other users find more easily contents in those specific languages.

4 RESULTS

The presumption of this work is that the rewarding mechanism has a certain influence on the body of posts, making users prefer to create contents regarding specific topics or cause other anomalies. Thereupon, one may stipulate that social networking effects (virality, shareability, friends/followers) might be changed significantly due to the emergence of financial reward. Since Steemit is similar to micro-blogging social networking platform, it will contain huge amounts of textual data in an unstructured format on a variety of topics, depending on each user and their interests. Accordingly, we adopt topic modelling to identify post topics in each of the three years and observe whether voting encouragement is present and find hidden topics and areas of interests for the users each year that are similar to each other. Topic modelling helps not only to specify the posting topics contained within the users' posting habits and according to their interests, but also help extract keywords which are related to the topics to develop an in-depth analysis.

The results of the topic modelling, therefore, might provide a helpful first look at understanding how users encourage others to vote for them and see the precise language utilization in this respect, irrespective of the subjectively defined published topic from the users, helping us analyze the underlying semantic structure of the posts.

We first establish the general word length of posts in the investigated period. Figure 3 depicts the word count distribution for posts in each of the years under investigation. One can observe that the average number of words per posts is lowest in 2019, compared to the two previous periods, but this can be possibly explained by the fact that we have collected data only from January through

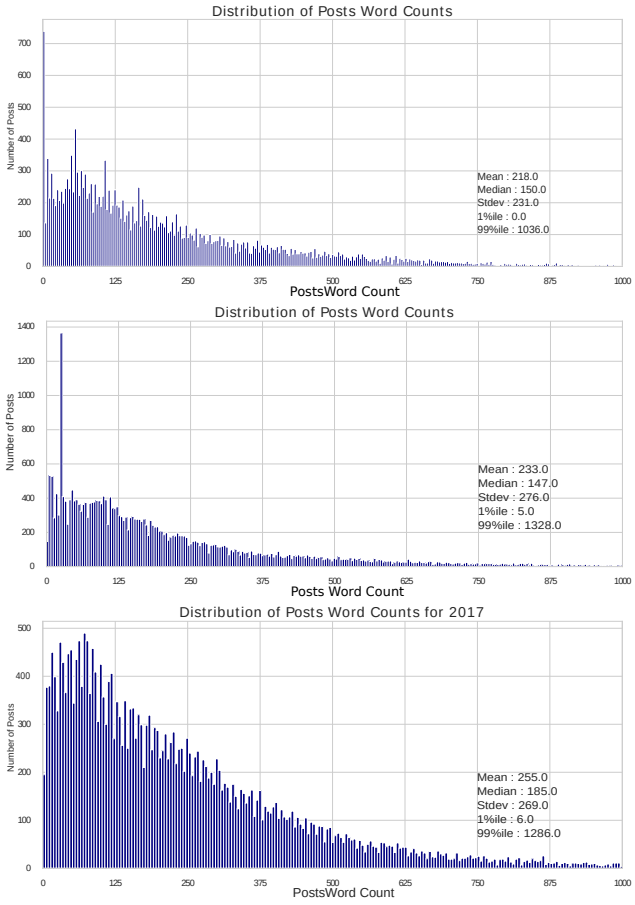

Figure 3: The three figures represent the word count distribution of the posts published for 2019 **(upper plot),** 2018 **(middle plot),** 2017 **(lower plot).**

September 2019, while we have data for the entire twelve months for 2017, 2018. Nevertheless, one can see a drop in the general word distribution from 2017 to 2018. Given that posts created in 2019 are much shorter, and because we do not have access to all the contents created in that year, for the rest of the paper we will only focus on the contents created either in 2017 or in 2018.

For the LDA analysis, the total number of topics (presented as a set of words) occurring in the collection of posts to be examined is 9. The topic naming convention was left as a topic together with the sequential number. The number of posts for each topic is presented in Figure 4, 2018 in the top plot, and 2017 in the bottom plot. In 2017 most of the posts belong to broad generic topics of mundane activities (Topics 1 and 3), but also two other topics emerge from the rest: a topic related to Steemit (Topic 5), and one related to food and nutrition (Topic 4). In 2018 the situation slightly shifted: while there is still a high number of posts about mundane activities (Topic 6), we also see an increased production in posts concerning Steemit (Topic 3) and cryptocurrencies in general (Topic 8). These results confirm that people are prone to talk about personal activities in social media, also in the case their information can be accessed to anyone on the Internet. Although among the more specific topics, the ones

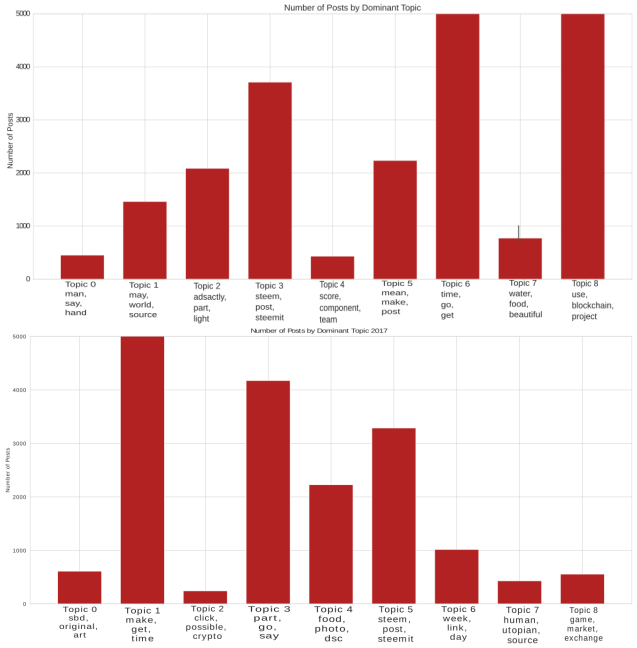

Figure 4: The two figures represent the number of posts per dominant LDA topic for 2018 **(middle plot), and** 2017 **(lower plot).**

related to cryptocurrencies, blogging, and Steemit, in particular, emerge over all the others. This is a very interesting and unique result, as we do not observe analogous behaviours for instance in Facebook or Twitter. The main reason behind this result surely lies in the fact that common people are eager to understand how the platform works without having the proper technical background to read the code of Steemit or do some reverse engineering. Therefore content producers are heavily encouraged to share with the other users insights about how the platform works.

In the rest of the section, we are going to inspect more closely the 9 topics for each year and the key words that define the topics, searching for possible anomalies. The results from the nine topics and the 10 words associated with each topic and their respective weight based on the vocabulary of the posts are depicted in Figure 5, the top plot for 2018, and the bottom plot for 2017. The separation by years was necessitated to develop a better understanding of the trends in topics, as well as to examine whether specific topics are emerging for each period. One should note that topic modelling will assign post fields and topics depending on the composition of the words that are ascribed to the topic.

As already emerged in the previous paragraph, in all three years are present topics directly related to Steemit, blogging, cryptocurrencies and blockchain in general (Topics 3, 5, and 8 in 2018, Topics 2, 5, and 8 in 2017) as well as some related to people's activity outside the platform. In particular, we see some topics related to generic mundane activities (Topics 1 and 3 in 2017, Topic 6 in 2018), but also some more specific topics, such as food/nutrition (Topic 7 in 2018, Topic 4 in 2017), sport (Topic 4 in 2018). It is worth to mention also Topic 0 in 2017, which at a first glance seems to be a very broad topic, but thanks to the word *welcome*, it is probably capturing all

the welcome posts of the users, that is, a post users are encouraged to make by the platform in which they present themselves, to make all the other users know something about the newcomers. Another interesting topic is Topic 6 in 2017, which contains the word *steemit*. While the other words do not necessarily seem to be linked with this one, it was very common for Steemit users to make daily/weekly contests or lotteries and award other users with small amounts of currency. Topic 2 in 2018 is very peculiar, indeed it mixes words like *art* and *story* with *witness* (so are called the Steem block creators). After further analyses, we discovered that the vast majority of the posts in this topic belong to a Steemit user called *ADSactly*, which is also the word with the highest weight of this topic. The account, which is probably not managed by a single person, posts a variety of contents about Steemit, Bitcoin and cryptocurrencies in general, as well as fiction, literature, arts, and so on.

Given their importance among all the topics, we will inspect more deeply the ones connected to Steemit. For the posts published in 2018, topic 3 stands out since it is related to both topics 2 and 3 from 2019. Here words about steemit, posting, commenting and upvoting have been outlined. Topic 8 is primarily outlined as one about blockchain, crypto-currencies, pricing, etc contained in topic 3. Finally, the situation is slightly bit different for 2017, where topic 5 relates to steem posts, voting, commenting and witnesses. Topic 8 at the same time contains information about the steem market, exchanges and price, which one can surmise relates to posts describing how Steemit works, including the exchange of currency, the blocks generated and trading that occurs. The results showed that irrespective of the year observed, a large number of posts have included in their body text about voting, steemit, steem, as seen from the topics containing information about *voting, steemit, post*.

5 DISCUSSION AND CONCLUSIONS

This research investigated the detection of the most common topics in Steemit, a platform that rewards content creators for their ability to engage many people. The intuition behind this study is that since the business model of the platform is different from its centralised counterparts, a sort of polarization and platform-specific topics would emerge. The analyses were carried out on the written part of the contents posted between January 2017 and December 2018, filtered by the English language, using Term frequency and Topic modelling techniques from NLP. The dataset was retrieved using the official API. The results point to the existence of some generic topics, along with more specific ones, throughout the three years of observations. The major finding is that a relevant part of the community is dedicated to producing contents about blockchain and cryptocurrencies in general, and on Steemit in particular. Thus, we can state that the business model adopted by Steemit heavily influences the contents created on the platform.

However, many other questions arise. One possible future work is to understand what drives the attention of the users in terms of voting to identify if some topics tend to be better rewarded than others. This is even more difficult if we consider that in the platform votes cast by users with more Steem Power count more than others. Moreover, as possible future work, we point out the detection of the strategies users adopt to increase the value of the contents they create to increase the rewards they get. Finally, we also plan to extend the work using Distributed Word Representations to have a more comprehensive understanding of the identified topics.

Acknowledgements This work is supported by the HELIOS H2020 project under grant agreement No 825585 and the ADAPT Centre, funded under the Science Foundation Ireland Research Centres Programme (Grant 13/RC/2106).

REFERENCES

[1] Reem ALRashdi and Simon O'Keefe. 2019. Deep Learning and Word Embeddings for Tweet Classification for Crisis Response. *arXiv preprint arXiv:1903.11024* (2019).
[2] Yoav Artzi, Patrick Pantel, and Michael Gamon. 2012. Predicting responses to microblog posts. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 602–606.
[3] Christopher Andrew Bail. 2016. Combining natural language processing and network analysis to examine how advocacy organizations stimulate conversation on social media. *Proceedings of the National Academy of Sciences* 113, 42 (2016), 11823–11828.
[4] S. Buchegger, D. Schioberg, L.H. Vu, and A. Datta. 2009. Implementing a P2P Social Network - Early Experiences and Insights from PeerSoN. In *Second ACM Workshop on Social Network Systems* (Nuremberg, Germany).
[5] Carole Cadwalladr and Emma Graham-Harrison. 2018. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The guardian* 17 (2018), 22.
[6] Rafael A Calvo, David N Milne, M Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering* 23, 5 (2017), 649–685.
[7] Yu-Ting Chang, Hueiju Yu, and Hsi-Peng Lu. 2015. Persuasive messages, popularity cohesion, and message diffusion in social media marketing. *Journal of Business Research* 68, 4 (2015), 777–782.
[8] Anne Cocos, Alexander G Fiks, and Aaron J Masino. 2017. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. *Journal of the American Medical Informatics Association* 24, 4 (2017), 813–821.
[9] L. A. Cutillo, R. Molva, and T. Strufe. 2009. Safebook: A privacy-preserving online social network leveraging on real-life trust. *Comm. Mag.* 47, 12 (2009), 94–101.
[10] Anwitaman Datta, Sonja Buchegger, Le-Hung Vu, Thorsten Strufe, and Krzysztof Rzadca. 2010. *Decentralized Online Social Networks*. Springer US, Boston, MA, 349–378.
[11] Derek Greene, Derek O'Callaghan, and Pádraig Cunningham. 2014. How many topics? stability analysis for topic models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 498–513.
[12] Alexander Gross and Dhiraj Murthy. 2014. Modeling virtual organizations with Latent Dirichlet Allocation: A case for natural language processing. *Neural networks* 58 (2014), 38–49.
[13] Barbara Guidi. 2020. When Blockchain meets Online Social Networks. *Pervasive and Mobile Computing* 62 (2020), 101131.
[14] Barbara Guidi, Tobias Amft, Andrea De Salve, Kalman Graffi, and Laura Ricci. 2016. DiDuSoNet: A P2P architecture for distributed Dunbar-based social networks. *Peer-to-Peer Networking and Applications* 9, 6 (01 Nov 2016), 1177–1194.
[15] Barbara Guidi, Marco Conti, Andrea Passarella, and Laura Ricci. 2018. Managing social contents in Decentralized Online Social Networks: A survey. *Online Social Networks and Media* 7 (2018), 12–29.
[16] Hee Jay Kang, Changhee Kim, and Kyungtae Kang. 2019. Analysis of the Trends in Biochemical Research Using Latent Dirichlet Allocation (LDA). *Processes* 7, 6 (2019), 379.
[17] Kristina G Kapanova and Velislava Stoykova. 2018. Social Network Analysis of "Clexa" Community Interaction Patterns. In *International Conference on Applied Physics, System Science and Computers*. Springer, 257–264.
[18] Amir Karami, Aryya Gangopadhyay, Bin Zhou, and Hadi Kharrazi. 2015. A fuzzy approach model for uncovering hidden latent semantic structure in medical text collections. *iConference 2015 Proceedings* (2015).
[19] Jiban Khuntia, Hang Sun, and Dobin Yim. 2016. Sharing news through social networks. *International Journal on Media Management* 18, 1 (2016), 59–74.
[20] Anna Sophie Kümpel, Veronika Karnowski, and Till Keyling. 2015. News sharing in social media: A review of current research on news sharing users, content, and networks. *Social media+ society* 1, 2 (2015), 2056305115610141.
[21] Hongmin Li, Doina Caragea, Cornelia Caragea, and Nic Herndon. 2018. Disaster response aided by tweet classification with a domain adaptation approach. *Journal of Contingencies and Crisis Management* 26, 1 (2018), 16–27.

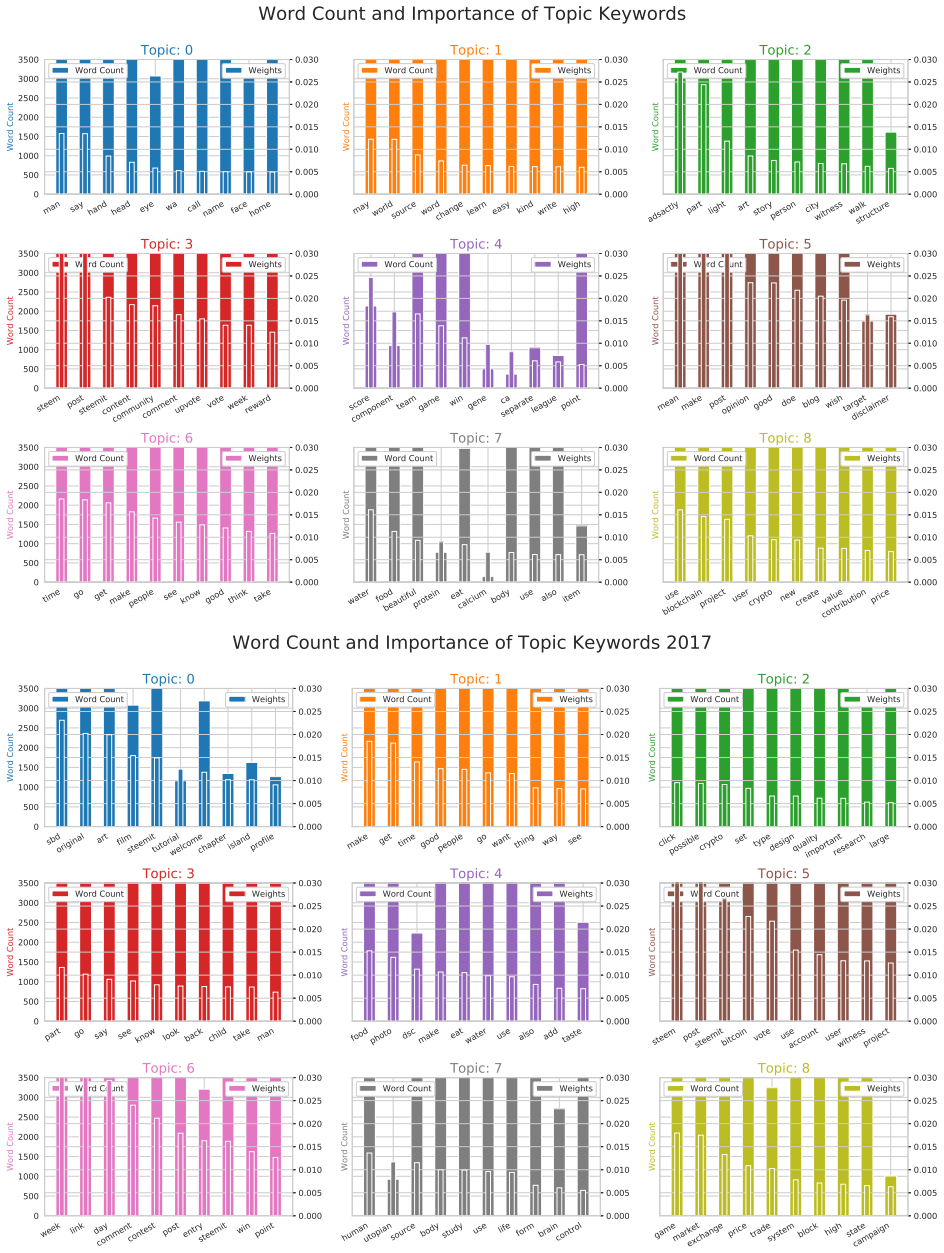

Figure 5: The two figures represent the LDA topics word importance for 2018 **(upper plot) and** 2017 **(lower plot).**

[22] Lydia Manikonda, Venkata Vamsikrishna Meduri, and Subbarao Kambhampati. 2016. Tweeting the mind and instagramming the heart: Exploring differentiated content sharing on social media. In *Tenth International AAAI Conference on Web and Social Media*.

[23] Arman Khadjeh Nassirtoussi, Saeed Aghabozorgi, Teh Ying Wah, and David Chek Ling Ngo. 2014. Text mining for market prediction: A systematic review. *Expert Systems with Applications* 41, 16 (2014), 7653–7670.

[24] Dat Tien Nguyen, Shafiq Joty, Muhammad Imran, Hassan Sajjad, and Prasenjit Mitra. 2016. Applications of online deep learning for crisis response using social media information. *arXiv preprint arXiv:1610.01030* (2016).

[25] Visar Shehu, Adrian Besimi, Urim Vejseli, and Douglas Jones. 2018. Towards Intelligent Disaster Response Systems. In *2018 ENTRENOVA Conference Proceedings*.

[26] Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring latent user properties from texts published in social media. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[27] Dong Wang, Boleslaw K Szymanski, Tarek Abdelzaher, Heng Ji, and Lance Kaplan. 2019. The age of social sensing. *Computer* 52, 1 (2019), 36–45.

[28] Jie Yin, Sarvnaz Karimi, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power. 2015. Using social media to enhance emergency situation awareness. In *Twenty-fourth international joint conference on artificial intelligence*.