Kurdish Studies
Jan 2024
Volume: 12, No: 1, pp. 3845-3856
ISSN: 2051-4883 (Print) | ISSN 2051-4891 (Online)
www.KurdishStudies.net

Received: October 2023 Accepted: December 2023
DOI: https://doi.org/10.58262/ks.v12i1.275

Text Analysis for Anomaly Detection and Fraud Mitigation in Social Media using R

Madhan N[1], Dheva Rajan S[2], Madhuri Jain[3]

Abstract

Online interactions are seriously threatened by the widespread fraudulent activity on social networking sites, which is typified using bogus identities and misleading viewpoints. The objective of the present work is to enhance Security Protocols and Digital Literacy on Social Media by proposing a robust anomaly detection model for social media. The present study proposes a comprehensive strategy that integrates technical, legal, and user-focused remedies such as content filtering, algorithmic transparency, and identity verification. When examining anomalies in social media data, Large Number of Rare Events models specifically, the left-skewed log-normal distribution become an invaluable statistical tool. The study uses Large Number of Rare Events models with equations related to vocabulary size and frequency class to solve the problems caused by uncommon events, such as cases of erroneous identities. Validation and application of LNRE models in social media data by utilizing text-based mathematical analysis to tackle fraud detection to be performed in the proposed work. The analysis of lexical diversity metrics offers language richness and a deeper understanding of the content under examination, aiding in the identification of focus areas, opinions, and discussions within the analysed text.

Keywords: *Gaussian, LNRE, lexical, social media, fraud, detection, anomalies*

Introduction

Social networking (SN) sites such as Facebook and Twitter have a great deal of power in influencing how people and organizations are seen. However, because people might pretend to be someone else and voice false thoughts, there is a chance that such a impact will lead to abnormalities. These types of fraud attempt to influence consumers' judgments by getting them to base decisions on incorrect information. The authenticity of online interactions is compromised by manipulation, which also jeopardizes the integrity of viewpoints expressed on digital platforms and may lead people or organizations to make decisions they otherwise would not have taken. There are serious repercussions associated with the widespread problem of fraudulent activity on SN sites, which includes the creation of false identities and false viewpoints. Such factors as deteriorating trust, harm to one's reputation, consumer misdirection, algorithmic biases, cybersecurity threats, influence on democracy, and false news can affect public opinion and decision-making, which may have an effect on democratic processes. Figure 1 gives the total loss for the year 2020 worldwide using various social media (SM) platforms.

[1] University of Technology and Applied Sciences Al Musannah, Oman
[2] University of Technology and Applied Sciences Al Musannah, Oman
[3] Banasthali Vidyapith, Rajasthan, India

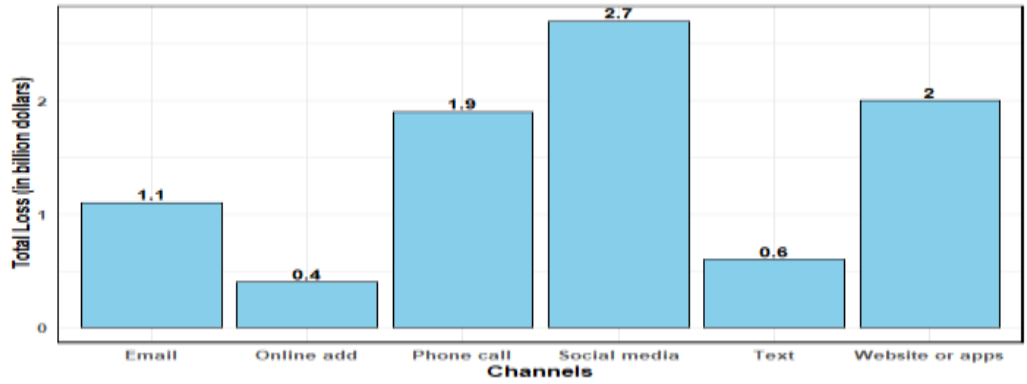

Fig:1 Total Loss Due to Fraud Using Various Social Media Platforms (Source: Https://Www.Ftc.Gov/).

Such SM fraud incurs a huge loss in economy worldwide. A glimpse of such economy loss is shown in figure 2.

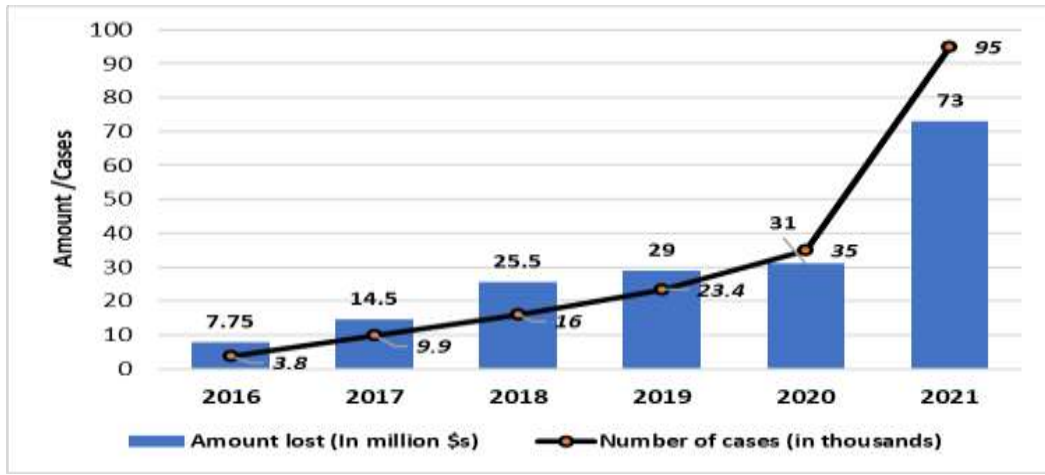

Fig:2 Economic Loss Increase Over the Years and Cases Registered in US (Source: Https://Www.Ftc.Gov/).

In order to counteract the consequences outlined above, platform administrators, users, and regulatory bodies must work diligently to improve security protocols, authenticate users, and encourage digital literacy in order to lessen the influence of fraudulent activity on SN sites. A multifaceted approach involving technological, regulatory, and user-centric solutions such as identify verification, content moderation, user reporting mechanisms, algorithmic transparency, educational initiatives, collaboration with authorities, continuous monitoring, user authentication technologies, and collaboration with other platforms is needed to address the problem of fake identities and misleading opinions on SN sites like Facebook and Twitter.

The present study aims to uncover anomalies like false opinions and phony profiles by proposing and demonstrating the usage of Large Number of Rare Events (LNRE) models for textual analysis of social media (SM) data. The goal of the research is to show how LNRE models may be applied to the said issue by giving example equations that connect textual characteristics like word frequencies and vocabulary size to the models. The study uses many existing methods to computationally examine the lexical variety of a randomly selected political

Facebook post and the comments that accompany it. Finding connections and patterns in the text that could point to possible abnormalities and irregularities is the aim. The overall goal of the present research is to demonstrate how useful LNRE models are for examining the lexical characteristics of SM text data as a useful method for maintaining reputations on social networks and spotting abnormalities.

Large Number of Rare Events Models

Large Number of Rare Events (LNRE) models are statistical models created to deal with circumstances in which a specific occurrence or result is uncommon but has important ramifications. LNRE models are intended to handle situations in which imbalances in the frequency of occurrences may cause conventional statistical approaches to be insufficient. These models are especially helpful in situations where some outcomes are rare but significant, like identifying and investigating rare events, like instances of false identities or misleading opinions, which can significantly affect the dynamics of social networks as a whole even though they are uncommon. They are also helpful in the analysis of rare diseases, extreme events, or anomalies in large datasets. The LRNE model is relevant because anomalies are examined in the present work.

Distributions that are left-skewed have an asymmetric shape and a longer tail, which suggests that the values are concentrated towards the upper end. Log-normal distribution is assumed by LNRE models, and LRNE fits in nicely with a lot of left-skewed datasets. Because it can accommodate both positive and zero values as well as different levels of skewness, it is appropriate for demonstrating such data. It is more acceptable to use the LNRE model since it accurately represents the relationship between variance and mean in left-skewed samples. The approach consists of constructing equations linked to vocabulary size, tokens, and frequency class of samples in order to demonstrate the applicability of LNRE models. LNRE models are useful for addressing anomalies in big datasets because they work well in situations where some events happen infrequently but can have a big impact. Additionally, the data will be investigated and shown using LNRE models. A feasible way of denoising the data by feature selection done by CoSelect (Tang & Liu, 2013) and it is based on SM posts.

Literature

For an instance, Twitter and Facebook, the two SM platforms control a sizable portion of the SN market. An entity's image, whether it be personal or corporate, is shaped by opinions expressed on these platforms. In the context of SN platforms, anomalies are defined as abnormalities or departures from the anticipated or typical behaviour. These might include different nature, influence on image, identification issues, and reputation management. Managing anomalies is essential for upholding a strong online reputation. In order to minimize any harm to one's reputation, prompt answers and remedial measures are important. As a result, abnormalities will inevitably occur. Customers are influenced by users who provide false identities and incorrect viewpoints. Customers are misled into making judgments by such deception that they otherwise would not have made.

SM data is mostly based on friendship and hence it contains excessive noise. Because SM data contains high dimensionality, the amount of noise in the data shoots up. (Tang & Liu, 2012) points to the fact that only a small part of the data is available only for mining. Rest is irrelevant and redundant. The users of SM can be classified into two categories – active content creators and passive content consumers (Liu et al., 2016).The anomaly problems were solved by the

fuzzy logic which is an extension of the set theory. Fuzzy can be used as a vital element in solving complex-problems which encompasses real-life problems (Drakulić et al., 2021) and helps in solving problems of uncertain sets (Allahviranloo & Pedrycz, 2020). Thus, combinatorics forms an integral part of the discussion. Qian et al., (2016) proposed the De-anonymizing social networks but by using knowledge graphs neither using texts nor by LNRE.

None of these actions is infallible, thus a mix of them is frequently required. Re-evaluate and modify approaches often to keep up with new and emerging scams on SM sites. The goal is to spot these fake accounts and take the appropriate action, such as banning, reporting, or deleting the user. The role of mathematical modelling becomes absolutely necessary to achieve the goal. It is possible to mitigate such problems using mathematical modelling, which also results in cost effectiveness. A model-oriented approach called LNRE is presented here to address the problem.

Proposed Methodology

Videos may be transcribed as text since the website content and comments are textual. The best model to investigate such scams is text-based analysis because it incorporated several languages and people's perceptions. The planned study would create a corpus of words by scraping data via open permission application programming interface (APIs) like Twitter or Facebook. In such case, Facebook is regarded as the data extractor. When it comes to sentiment lexical creation, the corpus approach has clear benefits over dictionary and human-based approaches. The corpus technique uses big datasets to automatically extract sentiment lexicons, in contrast to dictionary-based methods that might not grasp context-specific information and human-based methods that depend on subjective judgments and may be biased by an individual.

Modelling Equations

The link between the quantity of unique words and other factors, such as text length or genre, may be sculped via equations pertaining to vocabulary size. The distribution of words across various frequency classes and the relationship between that distribution and linguistic properties may be sculped using equations pertaining to the frequency class of the samples. Here, two equations will be proven to demonstrate the applicability of LNRE models.

One focuses on vocabulary size, while the other examines tokens (N) and sample frequency class. The total number of distinct words (or kinds) found in each text or corpus is known as the vocabulary size (V(N)). The term "Frequency Class of Samples" describes how words are arranged in a text or corpus according to how frequently they appear. Given a text of N tokens, let $V(N)$ be the different vocabularies and $V(m,N)$ be the frequency of words with m_n the largest frequency,

If the pdf of a Gamma (Φ, s) distribution of a variable p is defined as: $f(x) = \frac{\Phi e^{-\Phi x}(\Phi x)^{s-1}}{\theta(s)}$ and $E[p] = \frac{s}{\Phi}$. For $0 < s < 1$, because of the left skewness of the distribution Gamma distribution is the appropriate one to model the scenario.

Using the best fit concept between Gamma and inverse Gaussian, Gamma is ensured as a good fit for LNRE models.

Hence, there is a need to prove two equations as a proof of LNRE distributions:

1. $E[V(N)]=\frac{\Phi}{s}\left(1-\left(\frac{\Phi}{\Phi+N}\right)^{s}\right)$

2. $E[V(m,N)]=\frac{\Phi}{s}\,\frac{\theta(m+s)}{\theta(m+1)}\left(\frac{N}{N+\Phi}\right)^{m}\left(\frac{\Phi}{N+\Phi}\right)^{s}$

Proof: It is known to us that

$G(\omega)=s\int_{\omega}^{\infty}\frac{\Phi e^{-\Phi x}(\Phi\omega)^{s-1}}{\theta(s)}dx$, now $E[P]=\frac{s}{\Phi}$ or $\frac{\Phi}{s}=\frac{1}{E[P]}$

$$E[V(N)]=\int_{0}^{\infty}(1-e)^{-N\omega}\,dG(\omega)\qquad=\int_{0}^{\infty}(1-e)^{-N\omega}\,\frac{\Phi}{s}\,\frac{\Phi e^{-\Phi\omega}\Phi\omega^{s-1}}{\theta s}\,d\omega$$

$$=\frac{\Phi}{s}\left(1-\int_{0}^{\infty}\frac{\Phi e^{-(\Phi+N)\omega}(\Phi\omega)^{s-1}}{\theta(s)}\,d\omega\right)$$

$$=\frac{\Phi}{s}\left(\left(\frac{\Phi}{\Phi+N}\right)^{s}\int_{0}^{\infty}\frac{(\Phi+N)e^{(-\Phi+N)\omega(\Phi+N)^{s-1}\omega^{s-1}}}{\theta(s)}\,d\omega\right)$$

$=\frac{\Phi}{s}\left(1-\left(\frac{\Phi}{\Phi+N}\right)^{s}\right)$. Hence equation (1)

Likewise, for frequency function:

$$E[V(m,N)]=\int_{0}^{\infty}\frac{(N\omega)^{m}}{m!}\,e^{-N\omega}\,\frac{\Phi}{s}\,\frac{\Phi e^{-\Phi\omega}(\Phi\omega)^{s-1}}{\theta(s)}\,d\omega$$

$$=\frac{\Phi}{s}\int_{0}^{\infty}\frac{N^{m}}{m!}\,\frac{\Phi e^{-(N+\Phi)\omega}\Phi^{s-1}\omega^{m+s-1}}{\theta(m+s)}\,d\omega$$

$$=\frac{\Phi}{s}\,\frac{\theta(m+s)}{\theta(m+1)}\left(\frac{N}{N+\Phi}\right)^{m}\left(\frac{\Phi}{N+\Phi}\right)^{s}$$ Hence equation (2)

In the present work, it is examined the entire data using free opensource software R with packages RColorBrewer, tm, ggplot2, sos, ggpattern, wordcloud, and GGally. R studio the Integrated Development Environment (IDE) for R gives us the privilege to pull the SM posts or to upload the text file consists of post content in a large volume, R is highly suitable for such analysis comparatively with other available software. Mizumoto & Plonsky (2016) explained the suitability of R programming for linguistic analysis. An arbitrary political post is selected and analysed for the anomalies. The selected post is pulled from 10 news channels telecasted / published that via Facebook. The comments of those posts were pulled and analysed using

the proposed model to find various ratios for the analysis of lexical richness. It provides information on the variety of words used in a text as well as their quantity. It is used commentpicker.com to pull such comments of the particular post. Depending on the requirements of the task and the tools available, there are several ways to assess the lexical variety of a text or corpus. The word associates for the top 5 words were found. For an instance, the word nirmala has associated with modi (0.97), dmk(0.57), india (0.57), hindu(0.57) and hrnc (0.57). On the other hand, checking associations with the word bjp it is obtained the following: hindu (0.96), hrnc (0.96), temples (0.96), dmk (0.95), and india (0.96). The term association heatmap is given with 4 associated terms against 24 main terms in figure 3.

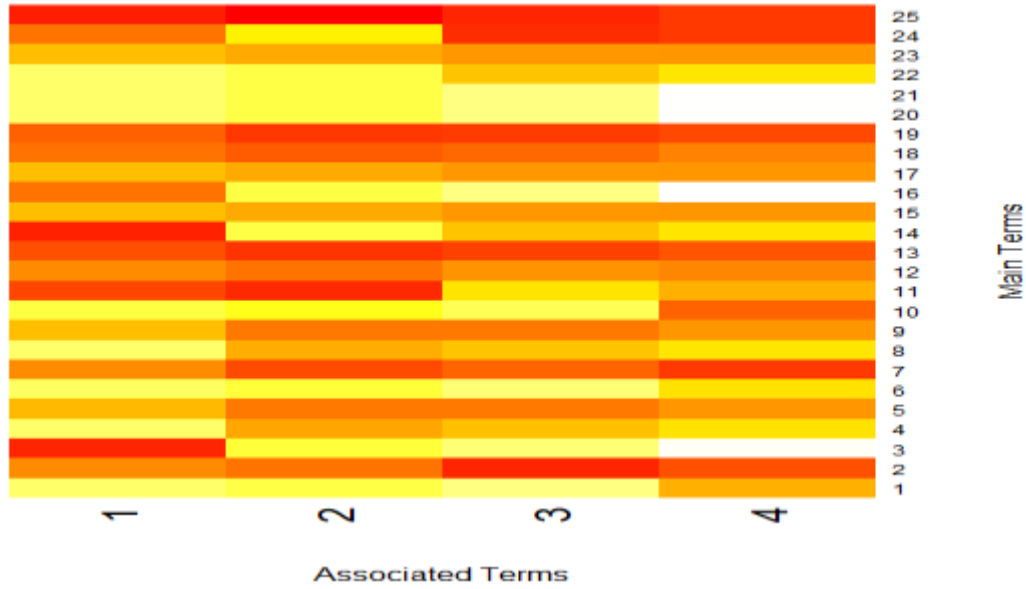

Fig:3 Term Association Heat Map.

Darker colors represent stronger positive correlations, indicating that the terms tend to co-occur frequently in documents. Lighter colors or blanks represent weaker or negative correlations, suggesting that the terms are less likely to cooccur. The summary statistics of the syuzhet_vector gives us -30.5 as the most negative sentiment score in text and 3.65 as the most positive sentiment score. Quartiles 1,2 equal to zero and mean equals 0.02, almost coinciding at 0. The text seems to have a distribution of sentiment scores around the neutral range, with a mean close to zero. There are both positive and negative sentiment expressions, but the overall sentiment leans towards neutrality. The summary of bing vector index having minimum value -4, mean equals -0.02, quartile 1,3 as 0, and maximum value equal to 4. The scale indicates that the texts are with high negative as well as positive sentimental words. Quartiles 1,2 and mean coincides at 0, shows a perfect coincidence at neutrality of words. On confirmation, the summary of afinn vector index having min value -10, quartile 1,2 and 3 coincides at 0 with the maximum value of 9. Hence ,75% of the values lies between -10 and 0, the remaining between 0 to 9 and ensure the sentimental results. Fourteen values of coefficients were determined and given in figure 4 and the remaining explained.

N/M index measures (as stated in Baayen, 2008) the ratio of the number of different words (N) to the total number of words (M). A value of 1.65 suggests a moderately diverse vocabulary. Lexical richness is measured by the TTR, which expresses the proportion of distinct words

(types) to all words (tokens) in the text. A TTR (Templin, 1957) of 0.61 indicates a moderate level of lexical diversity. It suggests that, on average, for every 100 words, there are 61 different words. A CTTR (Carroll, 1964) of 0.85 indicates a higher level of lexical diversity compared to the basic TTR. It suggests that, when accounting for text length, the vocabulary richness is relatively high. Guiraud's R or RTTR (Guiraud, 1960) is another measure of vocabulary richness, calculated by dividing the number of distinct words by the square root of the total number of words. A Guiraud's R of 57.2 suggests a relatively high vocabulary richness. It indicates that, on average, each root word appears in a context of about 57 words.

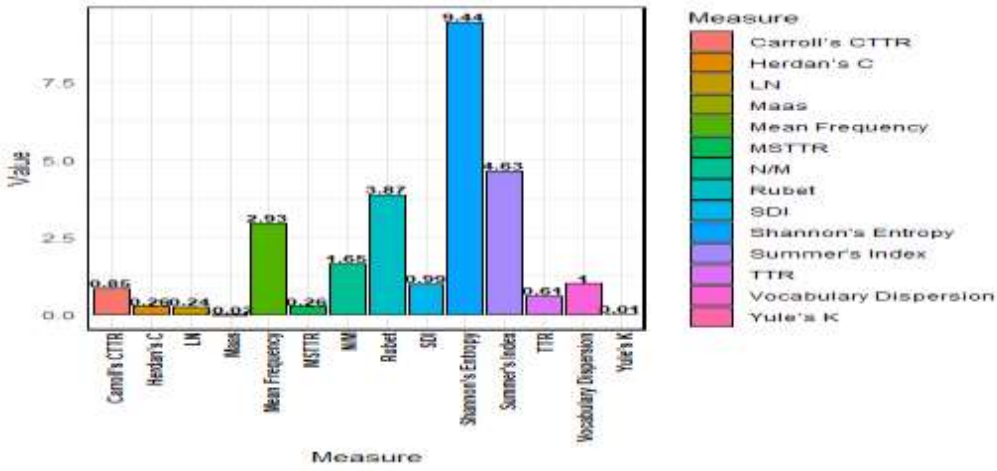

Fig:4 Values of Various Measures of Lexical Diversity.

A Herdan's C (Herdan, 1960, 1964) of 0.26 indicates a moderate level of vocabulary richness. It suggests that there is a reasonable variety of words in the text. TTR, word frequency, and text length are all combined to create composite metric called Summer's Index (1966) (as mentioned in Malvern et al., 2004). A Summer's Index of 4.63 is relatively high, indicating a diverse and rich vocabulary in the text. The higher the index, the richer the vocabulary. A value of 0.2 is common, and it might indicate a specific characteristic of lexical diversity in the text. Tuldava is a measure of lexical richness (Tuldava, 1993). A value of 0.2 suggests that there is 20 percent repetition of words in the text, indicating a moderate degree of lexical variety. A Rubet value (as mentioned in Malvern et al., 2004) of 3.87 indicates a relatively high level of vocabulary richness. An Honore's Richness Index (HRI) of 24.58 is relatively high, indicating a diverse and rich vocabulary in the text. (Honore, 1979)

Simpson's Diversity Index (SDI) calculates the likelihood that two text terms selected at random will differ from one another (Tweedie & Baayen, 1998). Maximum variety is indicated by a value of 1. The SDI in such instance is 1, indicating great lexical variety and the possibility that each pair of words in the text is unique. The degree of surprise or uncertainty involved in guessing the text's next word is measured by Shannon's Entropy (as mentioned in Zinin, 2016). The higher entropy (9.44) denotes increased unpredictability due to the significant lexical variety and a wide range of word usage. Dugast's S (Dugast, 1978) is a measure of lexical diversity, and a value of 1 indicates maximum diversity. In such context, it suggests a broad range of unique words in the text. Dugast's K measures the rate of change in vocabulary richness. A negative value like -197.65 suggests a decreasing trend in vocabulary richness, indicating that new words are being introduced at a decreasing rate.

Dugast's U (Dugast, 1978) is related to the vocabulary growth curve. Sichel's T (Zinin, 2016) is another measure of the vocabulary growth curve. A value close to 1 suggests a stable growth rate. A higher value indicates a more extensive vocabulary. The value of 186.2243 suggests a significant vocabulary size. Sichel's S (as mentioned in Zinin, 2016) measures the slope of the vocabulary growth curve. A higher value, such as 57.2, indicates a steep slope, suggesting rapid vocabulary growth. Yule's K (Yule, 1944) measures the distribution of word frequencies. A low value like 0.01 indicates a relatively even distribution of word frequencies in the text. Even though, the value of Yule's K is because of large corpus of words, as many words repeated.

Such words with frequency of frequency greater than 50 is given in figure 5. The horizontal axis of the bar graph shows specific terms that appear more than 50 times in the text. Each word's frequency, or the number of times it appears in the text, is shown on the vertical axis. One may determine which terms in the text are most frequently used and repeated by analysing the graph.

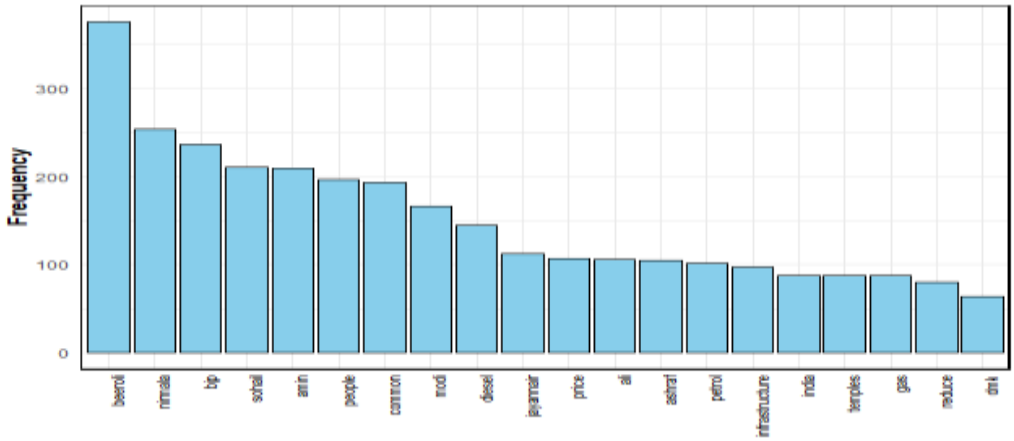

Fig:5 Frequency of Top 20 Words.

Word frequency concentration is evaluated using Herdan's C or sometimes log TTR. The fact that the result of 0.26 indicates a somewhat distributed distribution of word frequencies shows that the text does not mainly depend on a small number of terms. The observed growth in vocabulary is compared to the predicted growth using the Lukjanenkov & Nesitoj LN (Zinin, 2016) metric. With a score of 0.24, vocabulary growth is minimal compared to expectations. A Hypergeometric Distribution (HD-D) (P. M. McCarthy & Jarvis, 2010) value of 67.72 denotes a modest degree of dispersion, implying that the vocabulary was distributed fairly across the text.

Yule's Q (Yule, 1944) is another measure related to the distribution of word frequencies. A high value, such as 0.969, suggests an uneven distribution, with a concentration of a few frequently used words and a strong positive association between word pairs in the text. Such results implies that certain words tend to co-occur together more frequently than expected by chance. The high value indicates a notable degree of association or pattern in the distribution of word pairs in the analyzed text.

Herfindahl Hirschman Index (HHI) (Brezina et al., 2016) measures concentration. A low value like 0.0595 indicates a diverse distribution of word frequencies. Such concentration is addressed by figure 6, word cloud. Mean Frequency represents the average frequency of words in the text. A value of 2.93 suggests moderate frequency. By taking into account smaller textual units, the

MSTTR (Johnson, 1944) provides a measure of lexical variation at the segmental level. A score of 0.24 indicates more text variety, revealing differences within segments. Figure 6 gives the word cloud of the pulled Facebook post. Brunet's Index (Zinin, 2016) is a logarithmic measure of vocabulary richness. A higher value, such as 2635.83, indicates a larger vocabulary size in the analyzed text. Such high value suggests a diverse and extensive range of words used in the text.

Fig:6 Word Cloud.

When taken as a whole, the lexical diversity metrics point to a text that has a moderate to high degree of word use variety, a balanced distribution of word frequencies, and some degree of unpredictable word occurrence. The Rubet and Maas values, which are negative, point to certain texts in the pattern of vocabulary increase and word dominance.

With a Measure of Textual Lexical diversity (MTLD) (P. McCarthy M., 2005) of 939, the text appears to have a high degree of lexical variety, meaning that there are many different kinds of unique words in it. A higher concentration of word frequencies is reflected in HRI of 25, which suggests that a select few words are used more often in the text. With a rating of 57.69 for Guiraud's R, the book appears to have a broad vocabulary with a wide variety of original terms. Together, these metrics shed light on the text's vocabulary's concentration, dispersion, and richness.

These often-recurring words might be essential terminology, buzzwords, or colloquial jargon. The first 16 positions are occupied by terms like beeroli, nirmala, bjp, sohail, amin, people, common, modi, diesel, jayannair, price, ali, ashraf, fuel, and infrastructure. One can comprehend the viewpoint in the material by looking at the top repeated terms and the global cloud. Among the most often occurring terms are beeroli, amin, jayannair, and ashraf. As a result, greater attention to the profiles or the words occurred on the profiles' geniality is subject to investigation. Figure 7 gives the correlation between the words. Here, it is taken only the words with frequency greater than 150.

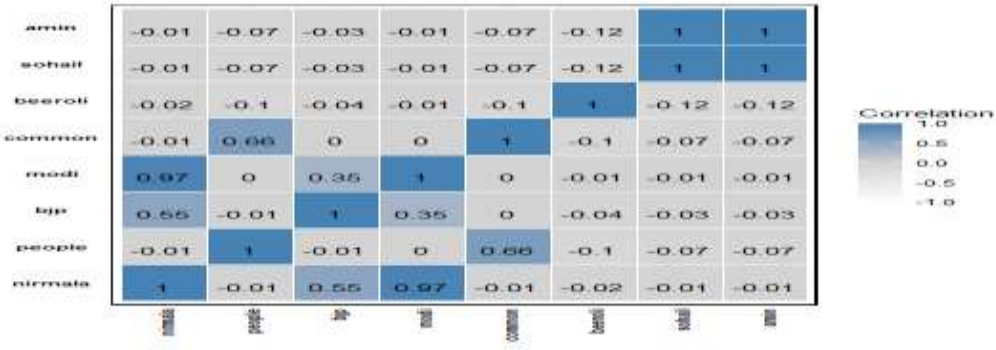

Fig:7 Correlation between the Words with Frequency Greater than 150.

At almost 0.97, the correlation between "nirmala" and "modi" is high, indicating a very strong positive association. Such association indicates that "nirmala" and "modi" occurrences typically rise or fall simultaneously. The data suggests a somewhat positive correlation of roughly 0.656 between the variable's "people" and "common". "Beeroli" and "common" have a correlation of -0.1006, which indicates a somewhat negative association. Here, it is to be noted that, the above values are correlation, and it does not imply causation. The other factors may influence the relationship between the words. Figure 8 illustrates the colocation network (Kim et al., 2023) of the frequent words found, shows where the words of the text document are cantered and gathered more.

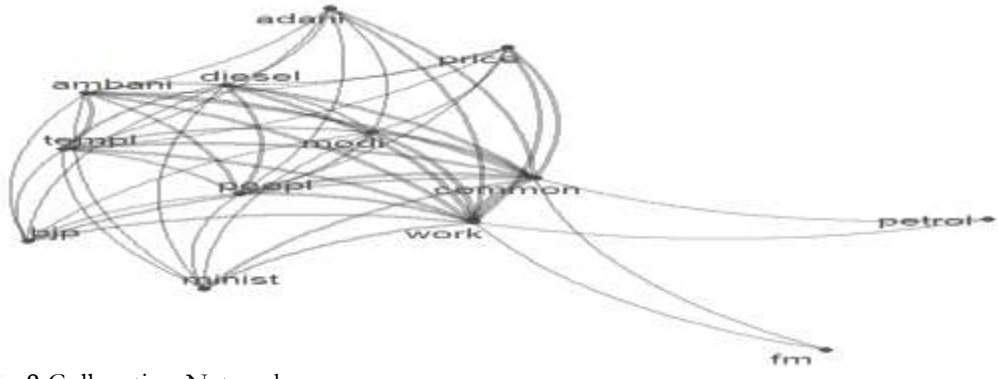

Fig:8 Collocation Network.

Each node in the figure represents a word in the text. Edges connect pairs of words that frequently co-occur, and the thickness or darkness of the edge may indicate the strength of the association. Thicker or darker edges represent stronger associations. The top features are likely the words with the strongest associations or frequent co-occurrences. The positioning of nodes in relation to each other might reveal clusters or groups of words that often appear together. Such diagrams are crucial in determining certain words that act as hubs connecting multiple other words.

Summary, Conclusion, and Future Work

It's critical to identify and act against fraudulent activity on SM, particularly with regard to phony identities and false beliefs. A careful approach is taken with the suggested methodology, which combines LNRE models with text-based mathematical analysis. The study concludes by navigating the complex terrain of SM fraud and highlighting the necessity of teamwork in order to improve security procedures and advance digital literacy. LNRE models provide a strong statistical foundation that is essential for tackling anomalies such as mistaken identities.

Diverse Vocabulary: The analysis of lexical diversity metrics indicates a text with a moderate to high degree of word use variety, suggesting a broad and diverse vocabulary.

Balanced Word Distribution: The metrics reveal a balanced distribution of word frequencies, with some degree of unpredictable word occurrence, enhancing the richness of the language used.

Analysis of Specific Words: Examination of specific words, their frequency, and associations provides valuable insights into the focus areas, opinions, and discussions within the analyzed text.

Correlation and Collocation: Correlation analysis between words and the collocation network diagram highlight the interrelationships between terms, allowing for a deeper understanding of how certain words co-occur and potentially influence each other.

The suggested methodology, which is based on a corpus-based strategy, shows how effective text-based analysis is in identifying frauds on SM. Important information may be gained from the correlation analysis of terms that appear often. The mathematical model ensures the authenticity whereas the comparison of various TTRs has been performed. It is also provided the word cloud and frequently repeated words were investigated. The correlation between the words gives the relationship between the words. The study's overall findings emphasize the significance of using preventative measures, such as ongoing monitoring and user authentication technology, to protect against the negative effects of fraudulent activity on SN sites.

The future development of the proposed model includes the usage of deep learning and other advanced Detection Algorithms like machine learning techniques to improve the LNRE model's anomaly detection and its capacity to identify subtle fraudulent trends. Proposing an AI model for providing an adaptive reaction and real-time monitoring system so that the LNRE model can quickly detect and handle new abnormalities in the ever-changing SM ecosystem.

References

1. Allahviranloo, T., & Pedrycz, W. (2020). Uncertain sets. *Soft Numerical Computing in Uncertain Dynamic Systems*, 13–65. https://doi.org/10.1016/B978-0-12-822855-5.00002-1
2. Baayen, R. (2008). *Analyzing linguistic data: A practical introduction to statistics using R* (1st ed., Vol. 1). Cambridge: Cambridge University.
3. Brezina, I., Pekár, J., Čičková, Z., & Reiff, M. (2016). Herfindahl–Hirschman index level of concentration values modification and analysis of their change. *Central European Journal of Operations Research*, *24*(1), 49–72. https://doi.org/10.1007/s10100-014-0350-y
4. Carroll, J. (1964). *Language and Thought. Englewood Cliffs* (46th ed.). Prentice-Hall Inc.
5. Drakulić, D., Takači, A., & Marić, M. (2021). The Use of Fuzzy Logic in Various Combinatorial Optimization Problems. *Studies in Computational Intelligence*, *973*, 137–153. https://doi.org/10.1007/978-3-030-72711-6_8/COVER/
6. Dugast, S. (1978). *Sur quoi se fonde la notion d'étendue théoretique du vocabulaire? 46*, 25–32.
7. Guiraud, P. (1960). *Problèmes et Méthodes de la Statistique Linguistique* (1st ed.). Paris: Presses universitaires de France.
8. Herdan, G. (1960). *Quantitative Linguistics*. Butterworth.
9. Honore, A. (1979). *Some simple measures of richness of vocabulary*. *7*.
10. Johnson, W. (1944). *Studies in language behavior: I. A program of research*. *56*, 1–15.
11. Kim, M., Oh, C. H., & Han, J. (2023). Colocation as network: Types and performance implications of structural positions in colocation network. *Journal of International Business Studies*. https://doi.org/10.1057/s41267-023-00629-8
12. Liu, H., Morstatter, F., Tang, J., & Zafarani, R. (2016). The good, the bad, and the ugly: Uncovering novel research opportunities in social media mining. *International Journal of Data Science and Analytics*, *1*(3–4), 137–143. https://doi.org/10.1007/S41060-016-0023-0/FIGURES/1
13. Malvern, D., D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment* (1st ed., Vol. 1). Houndmills, NH: Palgrave Macmillan.
14. McCarthy, P., M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Dissertation Abstracts International.

15. McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381
16. Mizumoto, A., & Plonsky, L. (2016). R as a Lingua Franca: Advantages of Using R for Quantitative Research in Applied Linguistics. *Applied Linguistics*, *37*(2), 284–291. https://doi.org/10.1093/applin/amv025
17. Qian, J., Li, X.-Y., Zhang, C., & Chen, L. (2016). De-anonymizing social networks and inferring private attributes using knowledge graphs. *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, 1–9. https://doi.org/10.1109/INFOCOM.2016.7524578
18. Tang, J., & Liu, H. (2012). Feature selection with linked data in social media. *Proceedings of the 12th SIAM International Conference on Data Mining, SDM 2012*, 118–128. https://doi.org/10.1137/1.9781611972825.11
19. Tang, J., & Liu, H. (2013). CoSelect: Feature selection with instance selection for social media data. *Proceedings of the 2013 SIAM International Conference on Data Mining, SDM 2013*, 695–703. https://doi.org/10.1137/1.9781611972832.77
20. Templin, M. (1957). *Certain language skills in children*. Minneapolis: University of Minnesota Press.
21. Tuldava, J. (1993). The statistical structure of a text and its readability. In *Quantitative text analysis* (1st ed., Vol. 1, pp. 215–227). Trier: Wissenschaftlicher Verlag Trier.
22. Tweedie, F. J., & Baayen, R. H. (1998). How Variable May a Constant be? *Computers and the Humanities*, *32*(5), 323–352. https://doi.org/10.1023/A:1001749303137
23. Yule, G. (1944). *The statistical study of literary vocabulary*. Cambridge University Press.
24. Zinin, S. (2016). Vocabulary richness of early Chinese texts: Macroanalysis of the Thirteen classics and the Zhuangzi. *Scholarly Reports of the Department of China of the Institute of Oriental Studies, Russian Academy of Sciences*, *XLVI*, 197–253.