MTNSA: Microblogging Text Network Sentiment Analysis (Incorporating Hashtags and Emojis)

Abdalrhman Alquaary[1],

[1] ALQUANIX, apo@alquanix.com, 0000-0002-8105-4101;

Abstract

In an era where social media platforms play a crucial role in shaping public discourse, microblogging data emerges as a vital resource for understanding complex social interactions. This paper introduces MTNSA (Microblogging Text Network Sentiment Analysis), a groundbreaking approach that harnesses the richness of social media communication by analyzing three separate categories: the relationships between words, relationships between hashtags, and relationships between emojis. MTNSA utilizes innovative techniques leveraging network theory to unravel the thoughts and opinions in microblogging environments, enriching itself by integrating sentiment analysis into this framework. This innovative method provides a comprehensive view of the sentiments associated with each node, offering deeper insights into the emotional nuances of online discourse. MTNSA's unique design enables its application across multilingual discourses, as it focuses on uncovering relationships between nodes, making it a versatile tool for global analysis in diverse linguistic contexts. The ability of MTNSA to blend nodes and emotional contexts into a unified analytical model presents a significant advancement in our understanding of digital communication patterns. It equips researchers, marketers, and policymakers with a robust tool to decode the intricate language of social media, contributing profoundly to our comprehension of how emotions and ideas are expressed and disseminated in the digital realm, thereby opening new frontiers for analysis in the dynamic landscape of social media.

Keywords: Network theory, text network analysis, sentiment snalysis, new media, social media.

1. INTRODUCTION

In the age of digital communication, microblogging platforms have emerged as vital spaces for public discourse, offering rich insights into collective sentiments and opinions. These platforms are vast and dynamic, which makes it significantly challenging to thoroughly analyze and interpret the text-based data they produce. This study introduces the Microblogging Text Network Sentiment Analysis, a novel approach designed to decode the complex web of interactions and sentiments inherent in microblogging environments.

At its core, MTNSA seeks to leverage the expansive textual data available on social media platforms, encompassing a diverse mix of words, hashtags, and emojis. The fundamental premise is that each element within a microblog carries potential insights into the sentiments and perspectives shared by its users. MTNSA is meticulously crafted to not only identify and categorize these elements but also to unravel the underlying sentiment and interconnections among them.

The methodology of MTNSA unfolds in a series of systematic steps: beginning with the data preparation phase, which involves cleaning and tokenization of microblogging texts. This is followed by node identification and preparation, where key textual elements are pinpointed and ranked based on their occurrence frequency. Subsequently, sentiment assignment to nodes is undertaken to contextualize the emotional undertones associated with these elements. The process culminates in the construction of a sentiment-informed text network, further visualized to represent the nuanced relationships and sentiment associations among the identified nodes.

This study is significant as it not only advances the methods of text analysis in microblogging contexts but also provides a comprehensive framework for understanding digital communication patterns and emotional undercurrents in online discourse. By bridging the gap between raw textual data and sentiment analysis defined by the user, MTNSA offers valuable insights for researchers, social media analysts, and digital marketers alike.

As the digital landscape continues to evolve, MTNSA stands as a testament to the power of innovative analytical approaches in harnessing the potential of big data. This study not only contributes to the academic discourse in the field of text analysis but also paves the way for practical applications in various sectors seeking to make sense of the ever-growing digital chatter.

2. LITERATURE OVERVIEW

Numerous studies have shown a keen interest in various aspects of social platforms, especially Twitter, focusing primarily on networks and relationships between users. One study delves into how user interactions on Twitter form complex social networks, reflecting the dynamics of information exchange[1]. Another research focuses on dissecting public discussions on significant topics like #SaudiWomenCanDrive, using social network analysis to identify key influencers and thematic communities[2]. Grandjean's work further contributes by emphasizing the role of linguistic groups in the digital humanities community[3]. Complementing these insights, Rosen's "Discrete Mathematics and Its Applications" offers a foundational theoretical framework for understanding complex network structures, vital for studying social media dynamics[4]. Together, these studies form a comprehensive picture of the multifaceted nature of social media networks.

In sentiment analysis, a range of techniques are utilized, with advanced transformer sentiment models being particularly noteworthy. For sentiment analysis in Arabic, the model developed by Inoue et al.[5] represents a significant advancement. In the context of English language analysis, Barbieri et al.[6] have established a comprehensive benchmark for tweet classification. Moreover, for a more global perspective, the multilingual sentiment analysis model by Barbieri, Espinosa Anke, and Camacho-Collados[7], known as the XLM-T model, demonstrates its effectiveness across various languages. While this paper leverages these specific models, the design of MTNSA methodology is versatile, offering users the flexibility to employ alternative techniques to suit their unique analysis needs.

Our research was undertaken to fill a unique niche in the realm of social media analytics. While numerous studies have focused on understanding the relationships between users on platforms like Twitter, our study shifts the focus to the relationships between words, hashtags, and emojis. We aim to delve deeper into the ideas and opinions expressed by individuals, exploring how these elements interconnect to form a broader narrative. This approach allows us to capture the essence of public discourse in a more nuanced way, examining the underlying ideas and thematic structures that define online conversations.

3. METHODOLOGY

This study introduces a novel approach, Microblogging Text Network Sentiment Analysis (MTNSA), designed to analyze and interpret the text itself in microblogging environments. Our methodology leverages the vast array of textual data available on social media platforms, encompassing a diverse mix of words, hashtags, and emojis. Each element within these microblogs carries potential insights into the sentiments and perspectives shared by users. The MTNSA process is meticulously crafted to not only identify and categorize these elements but also to understand the underlying sentiment and interconnections between them. By systematically dissecting and analyzing microblogging data, MTNSA aims to provide a comprehensive understanding of digital communication patterns and emotional undercurrents in online discourse. In the following sections, we will delve into the process of our algorithm, detailing each step in our approach.

Data Preparation

The MTNSA methodology initiates with a dataset comprising two columns: one for the text of each microblogging document and the other for its corresponding sentiment. The preliminary step in preparing this data involves a thorough cleaning process for the text column, especially the removal of stopwords, defined according to the criteria set by the user. Following this, we perform tokenization, a process where the cleaned texts are split into individual tokens (words, hashtags, emojis).

Equation (1) represents the initial dataset (Text Column), D. In this equation, D_1, D_2 ,… ,D_N signify the individual microblogging documents in the dataset, with N denoting the total number of documents.

$$\mathcal{D} = (D_1, D_2, \dots, D_N) \qquad (1)$$

Equation (2) defines the prepared dataset, $D_{prepared}$. It employs a union operation ($\bigcup$) over the entire range of documents in the dataset. For each individual document D_i, two functions are applied in sequence:

- **Cleaning Function (C)**: Applied to each document D_i, this function removes stopwords and other non-essential elements from the text, based on user-defined criteria.

- **Tokenization Function (T)**: After cleaning, the text is processed by the tokenization function *T*, which splits the cleaned text into individual tokens (like words, hashtags, emojis).

By applying Equation (2), each document in the dataset is transformed from its original format into a collection of tokens derived from the cleaned text. This tokenized format is more suitable for subsequent steps.

$$\mathcal{D}_{prepared} = \bigcup_{i=1}^{N} T\big(C(D_i)\big) \qquad (2)$$

Node Identification and Preparation

After the tokenization of nodes, the algorithm identifies and counts the nodes, prioritizing the top nodes based on their frequency of occurrence. It then selects a group of these top nodes for analysis, guided by a key hyperparameter, 'K'. This parameter specifies the number of top nodes to be considered, starting from the most frequent.

Equation (3) represents the tokenized form of the prepared dataset, where $T_1, T_2, \ldots, T_N$ are the sets of tokens from each document. Each T_i corresponds to the i^{th} document and contains the tokens resulting from the tokenization process.

$$\mathcal{D}_{prepared} = (T_1, T_2, \ldots, T_N) \qquad (3)$$

In equation (4), N is defined as the universal set of all unique tokens across all documents. It is formed by taking the union of the token sets from each document. This set represents all unique tokens that appear in the dataset.

$$\mathcal{N} = \bigcup_{i=1}^{N} T_i \qquad (4)$$

The frequency function $f(n)$ calculates the frequency of each token n in the set N. It counts the number of occurrences of token n across all documents, providing a measure of how often each token appears in the entire dataset.

$$f(n) = count(n, \mathcal{N}) \qquad (5)$$

Equation (6) defines F_K as a dictionary of tokens, selected based on their frequency. The selection function σ identifies the top 'K' tokens from the set N, with 'K' being a parameter that determines the number of tokens to be selected. This selection starts from the most frequented token. The dictionary F_K is made up of ordered pairs (n, f(n)), where 'n' is a token and 'f(n)' represents its frequency.

$$\mathcal{F}_{\mathcal{K}} = \{\big(n, f(n)\big) \mid n \in \sigma(\mathcal{N}, K)\} \qquad (6)$$

Sentiment Assignment to Nodes

In this stage, each node is assigned a sentiment score, calculated by the number of positive and negative occurrences in the microblogging texts. The sentiment scores from individual posts contribute to an overall sentiment score for each node. Nodes are categorized as positive or negative based on the majority sentiment. This dichotomous classification, represented by the $\psi(n)$ function, ensures that each node is distinctly categorized. $\psi(n)$ assigns 'Positive' if a node appears more frequently in positive contexts (pos(n) > neg(n)), and 'Negative' if it appears more in negative contexts (pos(n) < neg(n)) as it shown in equation (7). This method ensures nodes take a definitive stance, although it's important to note that natural sentiment can be applicable in some contexts as well.

$$\psi(n) = \begin{cases} 'Positive' \ if \ pos(n) > neg(n) \\ 'Negative' \ if \ pos(n) < neg(n) \end{cases} \qquad (7)$$

Equation 8 represents the enhanced dictionary FKsn, which includes each node n, its frequency $f(n)$, and its assigned sentiment $\psi(n)$. The set N_K refers to the selected top 'K' nodes. This dictionary enriches the frequency-based information with sentiment analysis, providing a comprehensive view of each node in terms of

both its occurrence and its sentiment association.

$$\mathcal{F}_{\mathcal{K}}^{\text{sn}} = \{(n, f(n), \psi(n)) \mid n \in \mathcal{N}_{\mathcal{K}}\} \quad (8)$$

Network Construction

In this phase, we construct a network for the top 'K' nodes, beginning with identifying and calculating the edges that connect these nodes across the corpus. Each interaction between the nodes is mapped, reflecting their connections, and laying the structural foundation of the network.

Relationships between nodes are established based on their co-occurrence in the same microblog. An edge is created in our network model whenever two or more selected 'K' nodes appear together in a text. The frequency of these connections is crucial, as it indicates the strength and significance of the relationships, and helps to understand the interconnectedness of the nodes within the microblogging texts.

Node Set (V): Equation 9 defines the set V, comprising all nodes n that are part of the top 'K' nodes, denoted by NK. V represents the collection of most significant nodes for network construction.

$$V = \{n \mid n \in \mathcal{N}_{\mathcal{K}}\} \quad (9)$$

Edge Set (E): This equation describes the set of edges *E*. An edge exists between two nodes *ni* and *nj* if they co-occur within the same microblogging text (represented by *Ti*). The condition *ni*=*nj* ensures that the edge is between two distinct nodes. The set *E* is formed by the union of all such pairs of nodes across all the texts in the dataset.

$$E = \bigcup_{i=1}^{N} \{ (n_i, n_j)\ n_i, n_j \in T_i, \\ n_i \neq n_j, and\ n_i, n_j \in V \} \quad (10)$$

Weight Function: This function calculates the strength of the connection between nodes *ni* and *nj*, considering their co-occurrences in the same text across the entire dataset. Since the graph is undirected, $w(n_i, n_j)=w(n_j, n_i)$, meaning the weight is the same regardless of the order of the nodes.

$$w(n_i, n_j) = \sum_{i=1}^{N} \text{count}(n_i, n_j, T_i) \quad (11)$$

Network Visualization

Following the "Network Construction", the visualization of the constructed network plays a pivotal role in understanding and interpreting the data. The visualization aspects are designed to provide a clear, intuitive representation of the network's characteristics and relationships. The key points of our visualization strategy are:

Node Size: The size of each node in the network directly corresponds to the frequency of the node within the corpus. Larger nodes indicate a higher frequency of occurrence, highlighting the most prominent or discussed nodes in the dataset.

Node Color: The color of each node is indicative of the sentiment associated with that node. There are three specific colors used to represent different sentiments:

- Green: This color is used for nodes associated with positive sentiments.
- Red: Nodes with predominantly negative sentiments are represented in red.
- Grey: Neutral sentiments are depicted using grey. This allows for a quick visual assessment of the overall sentiment landscape of the network.

Edge Thickness: The thickness of the edges between nodes signifies the strength of the relationship between them. Thicker edges imply a stronger or more frequent connection

between the nodes, indicating a higher level of interaction or co-occurrence in the texts.

Edge Color: In the MTNSA (Microblogging Text Network Sentiment Analysis), the color of the edges does not carry a specific analytical purpose. Instead, it serves a purely aesthetic or organizational role in the visualization, ensuring clarity and readability of the network diagram.
The extended network with visualization parameters can be represented as:.

$$G_{\text{viz}} = (V, E, s, c, t) \quad (12)$$

In this representation:

- V and E are the sets of nodes and edges, respectively, as previously defined.
- $s: V \rightarrow \mathbb{R}$ is the node size function, where larger sizes indicate higher frequency of occurrence.
- $c: V \rightarrow \{\text{Green, Red, Grey}\}$ is the node color function, indicating sentiment.
- $t: E \rightarrow \mathbb{R}$ is the edge thickness function, where thicker edges imply stronger connections.

An optional edge color function could be included for visual organization.

Important Note: It is crucial to emphasize that this algorithm is designed to operate on consistent entities. Users have the flexibility to generate three distinct types of network graphs: Words, Hashtags, or Emojis. Each of these graph types focuses on a specific aspect of the microblogging data. However, for a more comprehensive analysis, users also have the option to create a mixed graph that combines Words, Hashtags, and Emojis.

4. EXPERIMENTAL RESULTS

In our experimental analysis, we rigorously apply the Microblogging Text Network Sentiment Analysis (MTNSA) algorithm to distinct datasets, each offering a unique vantage point on public sentiment expressed via Twitter. By segmenting our analysis into hashtag, word, and emoji networks, we aim to dissect and understand the nuances of digital communication and sentiment in varied contexts. The hashtag network utilizes data from tweets bearing the "جائحة_كورونا" (COVID-19 pandemic) hashtags, shedding light on regional sentiments in Saudi Arabia during the pandemic. In contrast, the word network employs a separate dataset centered on the discussions about the TOGG car, Turkey's landmark electric vehicle, reflecting sentiments tied to national technological advancement. Finally, the emoji network is built from an entirely different dataset, comprising tweets from 10 unique Twitter accounts, to analyze the emotional undertones conveyed through these expressive symbols. This methodical partitioning allows us to explore the MTNSA algorithm's robustness in interpreting sentiments across diverse linguistic expressions and cultural contexts.

Hashtag Network Sentiment Analysis

The network depicted in Figure (1) showcases the centrality of the hashtag "جائحة كورونا" (COVID-19 pandemic), illuminating its pivotal role in the web of conversations. It stands out not only for its frequency but also for its strong connections to other significant nodes such as "كورونا" (corona) and "السعودية" (Saudi Arabia), indicating the depth of the pandemic's impact on public consciousness. The robust links between "جائحة كورونا" and other hashtags reflect the intensity of the discourse around the pandemic, as well as the public's sentiment towards various aspects of the crisis. The graph

distinctly portrays the duality of sentiment: the negative association with the pandemic itself and the more positive undertones related to the local responses and measures, encapsulating a society grappling with a global crisis yet finding pockets of positivity in communal and national efforts. This nuanced sentiment analysis underscores the MTNSA algorithm's adeptness in handling complex, multilingual datasets and extracting meaningful insights from a network of digital interactions. For this analysis, the 'K' value is set at 20 to determine the most significant hashtags for the network.

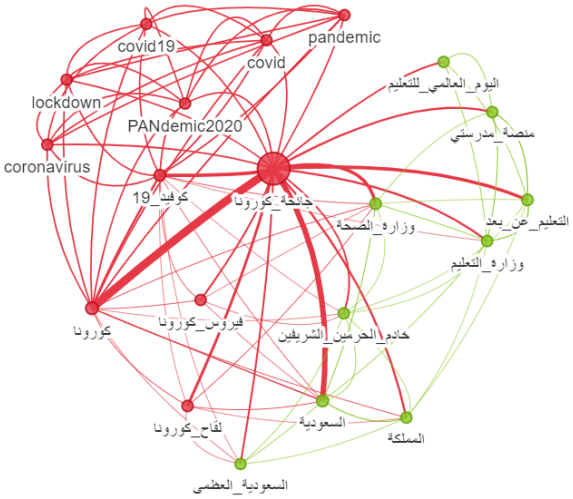

Figure 1 Hashtag Sentiment Network for COVID-19 (KSA)

Word Network Sentiment Analysis

Figure 2 presents a detailed sentiment network analysis of word associations linked to the TOGG hashtag, using a dataset that captures the public's response to the announcement of Turkey's new domestic car. The "TOGG" node is at the center, highlighting its significance with a strong connection to "TOGG2022," indicative of the focus on the vehicle's anticipated introduction in 2022. The overarching sentiment within the network is predominantly positive, showcasing a sense of national pride and excitement about the technological leap. For this analysis, a 'K' value of 25 was selected to identify the most pertinent words and hashtags for inclusion in the network.

However, the network also identifies a distinctly negative sentiment through the "BOGG" hashtag. This particular node, marked by its negative connotations, is used by a section of the populace to express dissatisfaction or criticism regarding the TOGG announcement. Analyzing the connections between "BOGG" and the other positive words and hashtags in the network provides a nuanced view of the public's varied reactions and the specific aspects of the TOGG announcement that did not resonate well with everyone. Understanding this dynamic is crucial for comprehending the full spectrum of public sentiment surrounding the introduction of TOGG.

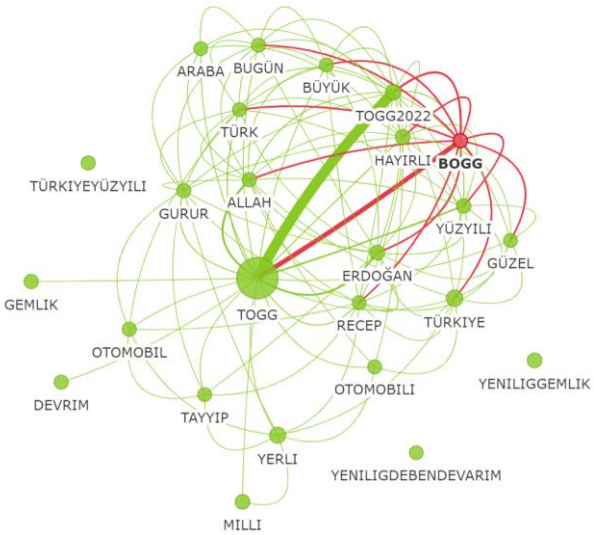

Figure 2 Word Sentiment Network for TOGG Tweets

Emoji Network Sentiment Analysis

Figure 3 illustrates the complex sentiment network of emojis utilized by 10 Saudi influencers on Twitter, composed of multiple clusters that likely correspond to various topics or contexts within the influencers' content. The intricate connectivity patterns suggest a rich

tapestry of discourse, with these key opinion leaders engaging on a broad range of subjects to interact with their audience. In constructing this network, a 'K' value of 39 was used to determine the most prominent emojis in the analysis, shaping the network's structure and highlighting the most influential symbols in the communication.

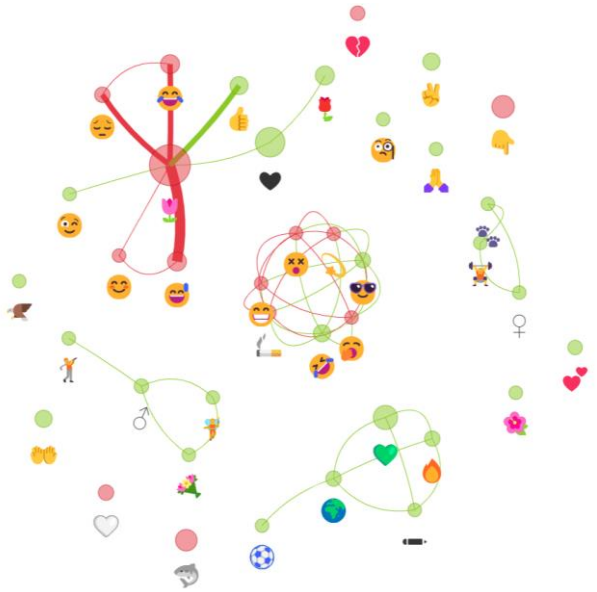

Figure 3 Emoji Sentiment Network for TOGG Tweets

The color-coded nodes provide an intuitive visualization of sentiment, with green nodes indicating a positive use of emojis. Remarkably, this positive sentiment is attributed to emojis typically associated with negative feelings, demonstrating the influencers' skill in recontextualizing emojis to fit the positive narratives of their messages. Conversely, emojis that are conventionally positive may appear in non-green nodes, suggesting a versatile use beyond their usual connotations. This dynamic use of emojis underscores the influencers' adeptness in crafting messages that resonate with their followers, reflecting the nuanced emotional engagement that these digital symbols can facilitate in social media communication.

Hashtag Network Analysis

When analyzing networks like the one in Figure 4, users are not confined strictly to our established methodology and may adapt features, provided they maintain logical consistency in their approach. This particular graph illustrates a network analysis of hashtags from the GEOSA_GOV_SA Twitter account's tweets, focusing on the connections between them. Nodes symbolize hashtags, and their size denotes the frequency of use, indicating their prominence in the discourse. Thickness of the lines represents how often hashtags are mentioned together. Colors in this graph are used only to differentiate elements and carry no analytical value. For this network, a 'K' value of 70 was selected to identify the top hashtags for the analysis.

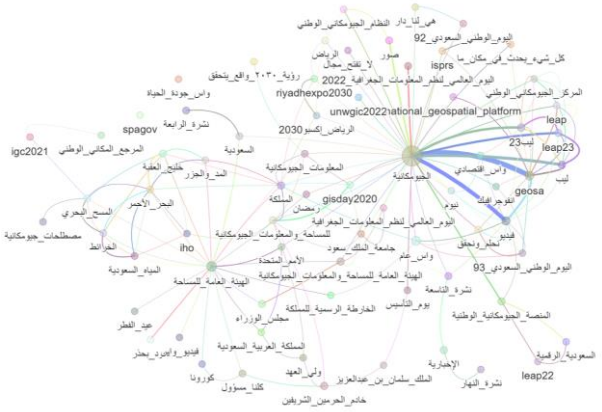

Figure 4 Hashtag Network for GEOSA_GOV_SA Account

5. DISCUSSION

In the discussion section of our study, we emphasize how our methodology, centered around the Microblogging Text Network Sentiment Analysis (MTNSA) algorithm, leverages pre-assigned sentiment values to construct networks. This approach enabled us to concentrate on the intricacies of relationships and interactions among various elements like hashtags, words, and emojis within these networks. By utilizing these pre-defined

sentiment values, we could more effectively showcase the sentiment of nodes and examine their impact on the network as a whole.

Our decision to use Twitter data was strategic, considering its widespread usage and the richness of its communicative elements. Twitter's structure, which prominently features hashtags, words, and emojis, offers a valuable dataset for this kind of analysis. It's important to note that while our analysis was rooted in Twitter data, the MTNSA algorithm's adaptable design allows for its application across different social media platforms, broadening its potential utility.

The use of multiple datasets in various languages was a deliberate choice to demonstrate the versatility and effectiveness of our method. In today's globalized social media landscape, the ability to process and interpret multilingual data is crucial. Our approach highlights the MTNSA algorithm's capacity to work effectively with diverse linguistic inputs, making it a robust tool for social media analysis.

Figure 4 exemplifies the flexibility of our approach. In this instance, we conducted a network analysis that focused on the relationships between hashtags in the tweets of the GEOSA_GOV_SA account. Unlike other parts of our study, the sentiments was not the primary focus here. Instead, the analysis aimed to understand the structural and relational dynamics of the data, demonstrating the MTNSA algorithm's ability to adapt to different analytical needs.

Acknowledgments
No Acknowledgement.

Funding
The author (s) has no received any financial support for the research, authorship or publication of this study.

REFERENCES

[1] I. Himelboim, M. A. Smith, L. Rainie, B. Shneiderman, "Classifying Twitter Topic-Networks Using Social Network Analysis," Social Media + Society, vol. 3(1), 2017.

[2] Z. Jastania, M. Ahtisham, R. Ayaz, K, Saeedi, "Using Social Network Analysis to Understand Public Discussions: The Case Study of #SaudiWomenCanDrive on Twitter," International Journal of Advanced Computer Science and Applications. vol. 11(2), 2020.

[3] M. Grandjean, "A social network analysis of Twitter: Mapping the digital humanities community," Digital Humanities, vol. 3, 2016.

[4] K. H. Rosen, "Graphs," Discrete Mathematics and Its Applications 7th ed, McGraw-Hill Education. 2012, ch. 10, pp. 641-735.

[5] G. Inoue, B. Alhafni, N. Baimukan, H. Bouamor, N. Habash, "The Interplay of Variant, Size, and Task Type in Arabic Pre-trained Language Models," Computational Approaches to Modeling Language (CAMeL) Lab New York University Abu Dhabi, 2021.

[6] F. Barbieri, J. Camacho-Collados, L. Neves, L. Espinosa-Anke, "TWEETEVAL: Unified Benchmark and Comparative Evaluation for Tweet Classification," Snap Inc. and Cardiff University, 2020.

[7] F. Barbieri, L. Espinosa Anke, J. Camacho-Collados, "XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond," Snap Inc. and Cardiff University, 2022.