IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 3, June 2026, pp. 2838~2848
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i3.pp2838-2848

Identification of areas of influence using a thematic modeling approach and belief function theory

Fatima-Zahrae Sifi[1], Wafae Sabbar[1], Amal El Mzabi[2]
[1]Laboratory of Machine Intelligence, Faculty of Sciences and Technology, Hassan II University, Mohammedia, Morocco
[2]Laboratory of Economic Performances and Logistics, Faculty of Law, Economic and Social Sciences, Hassan II University, Mohammedia, Morocco

Article Info

Article history:

Received Jul 14, 2025
Revised Mar 17, 2026
Accepted Apr 20, 2026

Keywords:

Belief function theory
Dempster-Shafer theory
Influencer analysis
Latent dirichlet allocation
Natural language processing

ABSTRACT

Social networks have become key platforms for the spread of information and the exchange of ideas. They change the way influencers interact and show influence over people. On platforms such as Twitter, Facebook, and Instagram, influence emerges over social interactions such as retweets, likes, mentions, comments, and shares. These activities have a major role in message amplification and opinion formation. Through the analysis of influencers' practices, it becomes possible to outline the areas where their impact is particularly significant. However, this task remains complex. Effective analysis requires robust methods that can incorporate diverse forms of social engagement and navigate the uncertainties posed by heterogeneous data. In this context, we propose a method to identify the domains of influence of a social media influencer. This approach combines thematic model latent dirichlet allocation (LDA) with belief function theory (BFT) to analyze social interactions and dominant topics of interest. The method examines indicators such as retweets, likes, and mentions and provides a robust framework to measure the influencer's impact across different domains. It thus offers precise tools for researchers, practitioners and decision-makers who aim to better understand these complex dynamics.

Corresponding Author:

Fatima-Zahrae Sifi
Laboratory of Machine Intelligence, Faculty of Sciences and Technology, Hassan II University
Mohammedia 28806 Morocco
Email: fatimazahrae.sifi@univh2c.ma

1. INTRODUCTION

In recent years, platforms like social media have become central to the circulation of information and public dialogue [1]. Their widespread use has changed the way we communicate. It provides users to share content instantly [2] and reach large diverse audiences [3]. In this online environment, influencers–individuals who command substantial attention and engagement [4]–exert considerable impact on opinions, trends and behaviors across various sectors. Platforms such as Twitter, Facebook, and Instagram facilitate this influence through interaction features like retweets, likes, mentions, comments, and shares [5]. These mechanisms contribute to the amplification and propagation of messages [6].

To understand how influence is exercised is a central challenge for researchers and practitioners. It is important to identify the domains of greatest influencer impact [7]. Traditional influence measurement techniques frequently ignore two critical aspects. The first is the thematic nuances of content [8]. The second is the intrinsic uncertainty and ambiguity in social engagement data [9]. This research areas underscores the need for advanced analytical models. These frameworks should combine machine learning-based topic modeling with robust methods for uncertainty management.

To address this gap, we propose a novel artificial intelligence (AI) approach. It uses latent dirichlet allocation (LDA) [10], a generative probabilistic model applied in machine learning for topic discovery. It also employs belief function theory (BFT), specifically the Dempster-Shafer framework [11], to adress uncertainty in influencer behavior analysis. This integration is our main contribution to AI. It combines a probabilistic generative model with evidence fusion techniques. The goal is to achieve a more robust identification of domain-specific influence. Our approach incorporates essential engagement indicators, such as retweets, likes, and mentions. It provides a multidimensional and context-sensitive perspective on influence. The proposed framework contributes both theoretically and practically. From a research perspective, it introduces an innovative combination of probabilistic modeling and belief fusion. This approach addresses uncertainty in social influence analysis. From an application perspective, it supports more effective strategies in marketing, opinion mining and digital communication. It also helps audiences better understand how influencers guide public discourse.

The structure of this paper is as follows. Section 2 reviews previous research in this domain. Section 3 describes the proposed method. It integrates LDA and BFT to identify and evaluate the domains of greatest influencer impact. Section 4 presents and discusses results from two case studies. Finally, section 5 concludes the study and outlines prospect directions for future research.

2. RELATED WORKS

Recent advances in AI have augmented the need for consistent methods to manage uncertainty. This is true for incomplete, ambiguous, or inconsistent data. Probabilistic frameworks [12] are common, but they frequently fail to represent ignorance or partial information clearly. This limitation has extended interest in approaches such as Dempster-Shafer theory (DST) [13].

The DST framework was introduced by Dempster and formalized by Shafer. It supports belief to be appointed to sets of hypotheses instead of exact probabilities. This makes it effective in the interest of both uncertainty and ignorance in areas such as sensor fusion [14], pattern recognition [15], and complex decision-making [16]. Its key components basic belief assignment (BBA), belief and plausibility measures and Dempster's rule provide a structured way to manage ambiguity and incompatible information [17].

Recent work has focused to reinforce the theoretical foundations of DST. One study [18] introduced new definitions of conditional belief functions and graphical models similar to Bayesian networks. This improved conditional reasoning. The authors of [19] examined the concept of "distinct" belief functions. They clarified the independence conditions needed to properly apply Dempster's combination rule. Dempster's original rule is elegant but performs poorly in high-conflict situations. To address this, alternative combination rules and entropy-based measures have been developed. One approach is a Rényi-style belief entropy [20]. It improves uncertainty measurement to support decision-making. In parallel, recent models that combine DST with topological evidence frameworks provide better conflict management. They approve belief computation to adjust in accordance with decision priorities [21].

Further research has integrated DST with topological models to improve conflict management. The author of [22] introduces flexible belief architectures. These architectures adjust belief calculations based on decision-making priorities. This facilitates DST to adapt to particular goals, such as reduce false positives or false negatives. It provides more focused solutions in complex decision-making contexts.

In decision theory, DST has influenced models that separate belief aggregation from decision-making. The transferable belief model (TBM) [23] defines two stages: the credal level beliefs and the pignistic level supports decision-making. A modern DST-based model addresses evidential uncertainty. It provides a structured manner to support rational decisions in situations needs quick judgment [24]. This extends DST's use to scenarios with incomplete or ambiguous information [25].

DST has been used broadly in practical fields. In sensor networks and the internet of things (IoT), it supports complex event analysis. It combines data from multiple sensors and controls uncertainty and conflict in dynamic environments [26]. DST is also used in deep learning for multimodal learning. It helps models integrate information from sources like images and text, principally when data is incomplete or concurrent. This is useful in tasks such as image descriptions and sentiment analysis [27].

In conclusion, DST continues to grow as a powerful mechanism for uncertainty management. It is useful where traditional probability models fall short. Its flexibility supports applications in inference, conflict resolution and decision-making. This demonstrates its rise role in AI and decision science.

3. METHOD

This study proposes a rigorous and automated approach to identify the specific domains where influencers have the greatest impact. It combines LDA [28] to extract the main topics from influencer-generated content, specifically tweets. It also uses BFT [23] to evaluate user engagement. Figure 1 illustrates the global

architecture of the proposed domain influence identification process. It shows each step from data collection to the identification of the main domain of influence. LDA identifies the main topics in influencers' discourse. Belief theory analyzes interactions such as retweets, mentions, and likes to measure confidence in each topic. The combination of these methods exposes both the content and how it resonates in the community. This provides a precise representation of influencers' areas of influence.

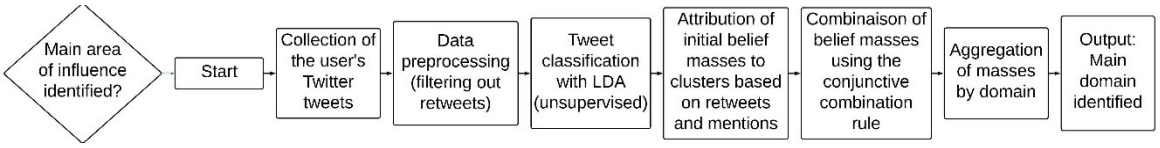


Figure 1. Global architecture for influence recognition within the domain

3.1. Formal framework of our method

Let Ω represent the set of potential topics or thematic areas relevant to a given influencer. For a specific influencer, Ω include topics such as {politics, technology, economy, and health}. We define a set of tweets $T_{initial} = \{t_1, t_2, \ldots, t_n\}$ posted within a time window $[t_{start}, t_{end}]$, where each t_i is a tweet published at a timestamp within this interval as in (1).

$$T_{initial} = \{t_i | t_{start} \leq timestamp(t_i) \leq t_{end}\} \quad (1)$$

For each tweet t_i we consider engagement metrics, where the number of retweets $r_i = Retweets(t_i)$, likes $l_i = Likes(t_i)$ and mentions $m_i = mentions(t_i)$. These metrics quantify both the influencer's engagement and the tweet's impact on the social network.

3.2. Thematic modeling using latent dirichlet allocation

To prepare the data for analysis, we first preprocess tweets by removing retweets, URLs, hashtags, punctuation and stopwords. The text is then lowercased, lemmatized and tokenized. This cleaned dataset forms the basis to use LDA to identify and classify primary topics in the tweets. Figure 2 illustrates the processing pipeline.

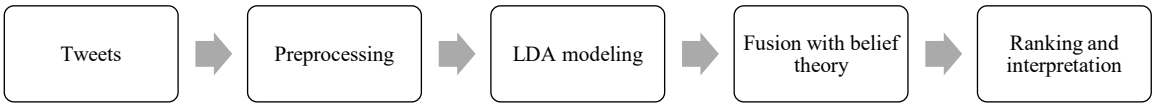


Figure 2. Sequential processing for the analysis and interpretation of Tweets

3.2.1. Data preprocessing

Preprocessing [29] starts by removing retweets to focus only on original content, as shown in (2).

$$T_{filtred} = \{t_i \in T_i / t_i \notin Retweet\} \quad (2)$$

Subsequent steps include :

- URL removal: $T_{noURL} = \{ t_i \in T_{filtred} \mid t_i \; contains \; no \; URLs \}$
- Removal of hashtags and special characters: $T_{noHashtag} = \{ t_i \in T_{noURL} \mid t_i \; contains \; no \; Hashtags \}$
- Punctuation removal: $T_{noPunctuation} = \{t_i \in T_{noHashtag} \mid t_i \; contains \; no \; Punctuation \}$
- Stopword removal: $T_{noStopwords} = \{t_i \in T_{noPunctuation} \mid t_i \; contains \; no \; Stopwords\}$
- Lowercasing: $T_{lower} = \{ t_i \in T_{noStopwords} \mid lower(t_i) \}$
- Lemmatization/stemming: $T_{lemmatized} = \{t_i \in T_{lower} \mid lem(t_i)\}$
- Tokenization: $T_{tokenized} = \{ t_i \in T_{lemmatized} \mid tokenize(t_i) \}$
- Noise reduction: $T = \{t_i \in T_{tokenized} | t_i \; is \; relevant \; for \; the \; analysis\}$

3.2.2. Tweet classification using latent dirichlet allocation

LDA is applied to discover latent topics within tweets [30]. LDA treats each tweet as a mixture of topics [31] and each topic as a distribution over the vocabulary. Each topic s is associated with a probability distribution over words as in (3).

$$\phi_s = \left(\phi_{s,1}, \phi_{s,2}, \ldots, \phi_{s,|V|}\right) \sim Dirichlet(\beta) \quad (3)$$

Where β controls topic-word sparsity and V refers to the vocabulary.

Each tweet $t \in T$ is modeled as a topic distribution as in (4).

$$\theta_t = \left(\theta_{t,1}, \theta_{t,2}, \ldots, \theta_{t,|\Omega|}\right) \sim Dirichlet(\alpha) \quad (4)$$

With α is used to control topic diversity within tweets. Tweets are thus classified by their dominant topics based on thresholded topic probabilities. This framework enables us to map tweets to meaningful topics. It provides information about influencers' topic preferences and discourse patterns.

3.3. Thematic influence and engagement metrics

Once LDA has been applied for topic modeling, we quantify audience impact by means of metrics. Thematic mass measures topic prevalence, while engagement masses (retweets, likes, and mentions) reflect audience interaction. To hold uncertainty in social interactions, we use BFT to model confidence in information sources. This enables a more robust evaluation of topic influence.

3.3.1. Thematic mass estimation

Once the LDA parameters are estimated, each tweet $t \in T$ is represented by a topic distribution $\theta_t = (\theta_t(s_1), \theta_t(s_2), \ldots, \theta_t(s_k))$, where $\theta_t(s)$ denotes the proportion of content in tweet t attributed to topic $s \in \Omega$, satisfying $\sum_{s \in \Omega} \theta_t(s) = 1$. To quantify the overall prominence of a given topic s within the corpus, we define the thematic mass $m_s(s)$. This measure aggregate measures the weighted presence of the topic across all tweets as in (5).

$$m_s(s) = \frac{\sum_{t \in T} \theta_t(s)}{\sum_{t \in T} \sum_{s' \in \Omega} \theta_t(s')} \quad (5)$$

This value captures the relative importance and prevalence of topic s in the dataset. Higher values indicate greater thematic dominance.

3.3.2. Retweet mass estimation

Let r_t represent the number of retweets associated with tweet t. To evaluate the level of audience engagement with each topic, we define the retweet mass $m_r(s)$ as in (6).

$$m_r(s) = \frac{\sum_{t \in T} \theta_t(s) r_t}{\sum_{t \in T} r_t} \quad (6)$$

This metric quantifies the relative share of retweet-driven attention allocated to topic s across the corpus. In contrast to thematic mass, retweet mass emphasizes the diffusion potential of topics. A high $m_r(s)$ indicates that topic s is frequently associated with tweets. It that generate substantial engagement, regardless of the total number of tweets.

3.3.3. Like mass estimation

Correspondingly, let l_t represent the number of likes received by tweet t. To quantify the level of positive audience engagement with each topic, we define the like mass $m_l(s)$ as in (7).

$$m_l(s) = \frac{\sum_{t \in T} \theta_t(s) l_t}{\sum_{t \in T} l_t} \quad (7)$$

This metric captures the emotional or affirmative response elicited by topic s across the corpus. A higher $m_l(s)$ value indicates that tweets associated with topic s tend to receive greater user appreciation. This suggests stronger resonance or agreement within the audience.

3.3.4. Mention mass estimation

The mention mass $m_m(s)$ quantifies the significance of topic s based on the frequency of its explicit references, such as hashtags or keywords, within the tweet corpus. For each tweet t, we combine its topic proportion $\theta_t(s)$ with the number of mentions m_t, and define the mention mass as in (8).

$$m_m(s) = \frac{\sum_{t \in T} \theta_t(s) m_t}{\sum_{t \in T} m_t} \quad (8)$$

This metric combines both the thematic relevance of topic s within individual tweets and the intensity of its explicit mention. Tweets with high $\theta_t(s)$ values and numerous mentions contribute more significantly to $m_m(s)$. Normalization by total mentions ensures valid comparisons across topics.

3.4. Mass fusion via sequential conjunctive fusion rule

To obtain a comprehensive measure of topic significance, we combine the mass functions—$m_s(s)$, $m_r(s)$, $m_l(s)$, and $m_m(s)$—through the sequential conjunctive fusion rule from belief theory. This method integrates evidence across different dimensions. It improves the alignment between thematic and engagement metrics. The step-by-step fusion process uses conflict coefficients to tackle uncertainty. This yields a dependable global influence score for each topic.

i) Fusion of thematic and retweet masses as in (9).

$$m_{sr}(s) = m_s(s) \oplus m_r(s) = \frac{1}{1-k_{sr}} \sum_{A \cap B = s} m_s(A).m_r(B) \quad (9)$$

With conflict coefficient as (10).

$$K_{sr} = \sum_{A \cap B = \emptyset} m_s(A).m_r(B) \quad (10)$$

ii) Fusion with like as in (11) and (12).

$$m_{srl}(s) = m_{sr}(s) \oplus m_l(s) = \frac{1}{1-k_{srl}} \sum_{A \cap B = s} m_{sr}(A).m_l(B) \quad (11)$$

$$K_{srl} = \sum_{A \cap B = \emptyset} m_{sr}(A).m_l(B) \quad (12)$$

iii) Fusion with mention mass as in (13) and (14).

$$m_{final}(s) = m_{srl}(s) \oplus m_m(s) = \frac{1}{1-k_{final}} \sum_{A \cap B = s} m_{srl}(A).m_m(B) \quad (13)$$

$$K_{final} = \sum_{A \cap B = \emptyset} m_{srl}(A).m_m(B) \quad (14)$$

At each fusion step, the conflict coefficient K quantifies the degree of inconsistency between the different mass sources. This coefficient enables the rule to adjust conflict evidence. It ensures a balanced integration of the metrics. The final mass $m_{final}(s)$, reflects the aggregated influence of topic s. The topic with the highest $m_{final}(s)$ is identified as the dominant influence within the corpus, reflecting its overall prominence across all dimensions of engagement.

3.5. Pignistic probability

To rank topics by their influence, we compute the pignistic probability ($BetP(s)$), which converts the final belief mass m_{final} into a probability distribution over individual topic. The pignistic probability is defined as in (15).

$$BetP(s) = \sum_{A \ni s} \frac{m_{final}(A)}{|A|} \quad (15)$$

Where $A \ni s$ denotes all subsets containing topic s and $| A |$ represents the cardinality of subset A.

This measure reflects the expected belief a rational agent would assign to topic s when making decisions under uncertainty. A higher $BetP(s)$ indicates a greater inferred influence of topic s. Therefore, the dominant influence domain is determined as the topic with the maximum $BetP(s)$.

4. RESULTS AND DISCUSSION

In this section, it is explained the results of research and at the same time is given the comprehensive discussion. Results can be presented in figures, graphs, tables, and others that make the reader understand easily [14], [15]. The discussion can be made in several sub-sections.

This study utilizes two Twitter datasets from influential figures—Jack Dorsey and Cristiano Ronaldo—enabling a comparative analysis of communication styles and social media engagement.

i) Jack Dorsey dataset: due to application programming interface (API) restrictions following Twitter's transition to "X" in early 2023, we manually collected 400 tweets from Dorsey's account. The dataset consists of original tweets, replies and retweets selected to ensure both temporal and thematic variety. Each tweet is accompanied by structured metadata, which includes ID, date, text, likes, and retweets. All data are stored in comma-separated values (CSV) format. This manual collection method helped maintain the accuracy and relevance of the data. The dataset is used for thematic analysis with LDA applied to identify topics in Dorsey's digital discourse. Additionally, it also supports the measurement of audience engagement.

ii) Cristiano Ronaldo dataset: collected via the Twitter API through Tweepy in 2019, this dataset [32] contains 2,507 original tweets by Cristiano Ronaldo over six years. Retweets were excluded to focus on user-generated content. Each tweet includes metadata such as timestamp, text, language, likes, and retweets stored in CSV format. The dataset reflects diverse communicative styles, from promotional to personal messages. It serves to explore digital identity construction and audience interaction with LDA-based topic modeling and engagement analysis applied to the data.

4.1. Results from the Jack Dorsey dataset

A comprehensive thematic analysis was conducted on tweets from Jack Dorsey, co-founder of Twitter and key figure in technology and cryptocurrency. The objective was to identify dominant topics and evaluate their importance. Probabilistic topic modeling LDA was combined with BFT for this task. The preprocessing steps involved cleaning the text, removing stopwords, lemmatizing and normalizing temporal and numeric entities. Tweets were tokenized and bigrams were incorporated to improve the analysis before vectorization.

The LDA model was trained on the preprocessed corpus using Gensim. The number of topics was set to 7, based on coherence optimization. The Dirichlet priors were configured as α = 'auto' (optimized per document) and β =0.01. This configuration ensured balanced topic sparsity and interpretability.

Model performance was evaluated using perplexity and topic coherence metrics. The model achieved a perplexity score of -7.35 and a C_v coherence score of 0.67. These results indicate strong internal consistency among topics. Additionally, the UMass coherence value of -20.21 reflects a stable distribution of semantically coherent topic-word associations across runs.

Table 1 presents the key topics identified in the Jack Dorsey dataset along with their dominant keywords. Each topic was allocated a label based on its dominant keywords. The most representative topics included bitcoin, money, bank, data, protocol, and finance.

Table 1. Key topics and dominant keywords in the Jack Dorsey dataset

Topic	Dominant keywords
Bitcoin	bitcoin, tool, control, global, declaration, cash…
Money	money, value, inflation, privacy, saving…
Bank	bank, crisis, infrastructure, money…
Data	data, content, content, technology…
Protocol	protocol, platform, signal, corruption…
Finance	finance, control, world, permission, bank…

The "bitcoin" topic stands out with a thematic mass of 0.763. It surpasses other topics such as money (0.063), data (0.053) and finance (0.038) and demonstrates a clear focus on cryptocurrency. To evaluate audience impact, we analyzed the engagement associated with each topic and we compared retweet and like distributions. Table 2 presents the thematic mass and audience engagement metrics (retweets and likes) for each identified topic in the Jack Dorsey dataset.

Table 2. Topic and engagement metrics in the Jack Dorsey dataset

Topic	Thematic mass	Retweet mass	Like mass
Bitcoin	0.763	0.525	0.508
Money	0.063	0.036	0.034
Protocol	0.048	0.031	0.026
Data	0.053	0.026	0.037
Bank	0.035	0.024	0.026
Finance	0.038	0.026	0.024

The "bitcoin" topic is the most prominent and attracts the largest audience, with higher content focus and audience interaction than other topics. To combine information from semantic, temporal and

engagement dimensions, we used a sequential mass fusion method based on BFT. This approach consolidates thematic importance and accounts for uncertainty across metrics.

After fusion, the "bitcoin" topic had a notably high fused mass of 0.300. All other topics showed near-zero values. The BetP confirmed bitcoin's dominance, with a value of 0.285, higher than other topics (≈0.143). This result reinforces its central role in the digital discourse. Table 3 presents the pignistic probability (BetP) values after sequential fusion for each topic.

Table 3. Pignistic score comparison

Topic	BetP (sequential fusion)
Bitcoin	0.2855657082
Money	0.1429114793
Protocol	0.1428837286
Data	0.1428925255
Bank	0.1428727251
Finance	0.1428738334

The results show that Jack Dorsey's Twitter discourse focuses on cryptocurrency. Bitcoin serves as a key element of his digital presence. Other topics appear, but their lower BetP scores indicate secondary importance. This reflects a focused communication strategy that gives priority crypto-related narratives, despite Dorsey's roots in web technologies.

To validate the topic modeling results obtained with LDA, the method uses transformer-based embeddings for representation of topics (BERTopic) combined with clustering and dimensionality reduction. This model enables a more precise detection of semantically coherent topics, especially for short, context-rich messages such as tweets. After preprocessing the same corpus, BERTopic automatically identified six dominant topics. They correspond to the same thematic families as LDA: bitcoin, money, finance, protocol, data, and bank. Table 4 presents the BERTopic results and include representative keywords, topic frequency, and topic probability for each identified topic.

Table 4. BERTopic results on Jack Dorsey dataset

Topic	Representative keywords	Frequency (%)	Topic probability
Bitcoin	bitcoin, money, system, trustless, network	41.6	0.72
Finance	bank, system, central, failure, traditional	19.4	0.61
Protocol	protocol, platform, code, layer, privacy	13.2	0.58
Data	data, content, technology, user, control	11.8	0.55
Money	money, inflation, value, saving, currency	8.9	0.49
Bank	bank, system, central, failure, traditional	5.1	0.45

A cross-comparison of both methods (LDA vs. BERTopic) indicates that they converge on the same key topics, particularly bitcoin, finance, and protocols. This result demonstrates the robustness of the discourse structure. While LDA quantifies topic distribution based on word frequency, BERTopic clusters semantically related tweets according to meaning and context. This creates refined distinctions within the same conceptual families. Both analyses confirm bitcoin's thematic centrality, the recurrence of financial and technological issues and the limited presence of secondary topics such as data and bank. Rather than opposing both approaches, their convergence strengthens the interpretative validity of the findings: Dorsey's Twitter activity is centered on cryptocurrency ideology but also interlinks technology, decentralization and financial independence that reflect a coherent digital identity across methods.

4.2. Results from the Cristiano Ronaldo dataset

A thematic analysis of tweets from Cristiano Ronaldo, a globally influential football figure, was performed to identify recurring content topics and assess their relative prominence. The analysis combined LDA with BFT and also implemented in Gensim to consolidate findings from both semantic structure and audience engagement. The model was configured with 12 topics, chosen to balance granularity and interpretability, with α automatically optimized per document and β set to 0.01 to encourage sparse topic-word distributions. Model evaluation yielded a perplexity of -10.39, a C_v coherence score of 0.60 and a UMass coherence of -17.50, indicating reasonable semantic coherence across topics. LDA identified four dominant topics based on semantic proximity. Table 5 lists the main topics identified in the Cristiano Ronaldo dataset along with their dominant keywords.

Table 5. Key topics and dominant keywords in the Cristiano Ronaldo dataset

Topic	Dominant keywords
Game	game, match, worldcup, portugal...
Madrid	madrid, goal, team, favorite, match...
Thanks	fan, photo, support, video...
Portugal	Portugal, game, match, parabens…

The Madrid topic emerged as the most prominent, with a thematic mass of 0.331, followed by game (0.254), thanks (0.221), and Portugal (0.195). These results show a strong association with Real Madrid, a central element of Ronaldo's public identity. Table 6 presents the thematic mass alongside retweet and like engagement metrics for each topic.

Table 6. Topic and engagement metrics in the Cristiano Ronaldo dataset

Topic	Thematic mass	Retweet mass	Like mass
Madrid	0.331	0.257	0.215
Game	0.254	0.211	0.210
Thanks	0.221	0.199	0.216
Portugal	0.195	0.140	0.152

To refine the analysis, social engagement metrics was integrated. The Madrid topic guided into audience reception with the highest retweets and likes. Game, thanks, and Portugal also had significant engagement but were secondary. These results underscore Real Madrid's strategic prominence in Ronaldo's digital communication, which confirms a consistent identity tied to club legacy and fan interaction.

A topic weight based on audience engagement, measured through "likes" and "retweets" was also introduced. This cross-evaluation confirmed the continued dominance of the Madrid topic with the highest engagement values (retweet mass: 0.257, like mass: 0.215). In contrast, game (retweets: 0.211; likes: 0.210), thanks (retweets: 0.199; likes: 0.216), and Portugal (retweets: 0.140; likes: 0.152) had lower levels of attention but remained aligned with the overall thematic framework. To consolidate these interpretations, a sequential fusion strategy based on BFT was applied. Table 7 presents the BetP values obtained from the sequential fusion process.

Table 7. Pignistic score comparison

Topic	BetP (sequential fusion)
Madrid	0.2848
Game	0.2521
Thanks	0.2438
Portugal	0.2191

The pignistic probability for the Madrid topic was 0.284, much higher than other topics. This confirms its central role in Ronaldo's digital communication. Real Madrid references are both thematically dominant and strongly connected with his audience. They reflect a consistent identity built around club legacy and personal brand.

The LDA combined with BFT shows the Madrid topic as central to Cristiano Ronaldo's digital discourse and gives structure to his Twitter communication. It reflects his history with Real Madrid and his strategy to support his public image in athletic achievements. Unlike other public figures who diversify into entrepreneurship, politics or philanthropy, Ronaldo keeps a sports-centered narrative. Secondary topics like game, thanks, and Portugal support engagement and community but stay aligned with his football identity. They maintain a focused brand strategy rooted in his athletic legacy. Beyond the LDA framework, a complementary experiment with BERTopic was conducted to uncover semantic layers within Cristiano Ronaldo's tweets. Table 8 summarizes the BERTopic topics extracted from the Cristiano Ronaldo dataset. It included representative keywords, frequencies and topic probabilities.

The model produced five cohesive thematic clusters; each one represents a key dimension of Ronaldo's communication. The aim was not to replace LDA but to compare the thematic segmentation of both models. This ensures interpretive consistency across computational approaches. Both LDA and BERTopic converge on the same core thematic areas. Madrid, game, and Portugal are the main axes of Ronaldo's digital discourse. While LDA describes the probabilistic structure of topics based on word frequency, BERTopic reinforces semantic cohesion. It groups related tweets under broader clusters.

This complementary perspective confirms that Ronaldo's communication strategy is centered in club identity and sports performance. It also integrates personal and emotional narratives (thanks and motivation) that strengthen fan proximity and engagement. The convergence of results from both models supports the robustness and consistency of the thematic interpretation. Cristiano Ronaldo's Twitter discourse, therefore, reflects a coherent and multidimensional narrative in which professional excellence and personal authenticity coexist to sustain audience connection.

Table 8. BERTopic topics extracted from the Cristiano Ronaldo dataset

Topic	Representative keywords	Frequency (%)	Topic probability
Madrid	madrid, goal, champions, team, club	33.7	0.70
Game	match, win, worldcup, training, victory	27.1	0.65
Portugal	portugal, selecao, fans, nation, pride	18.4	0.61
Thanks	thank, support, love, photo, family	12.3	0.57
Motivation	proud, dream, together, moment, home	8.5	0.54

5. CONCLUSION

This study presents a framework that combines LDA with BFT to analyze social media discourse. Topics are modeled probabilistically and audience engagement, such as likes and retweets, are incorporated. This approach provides a clearer view of what drives communication among influencers. The key advantage of this methodology resides in its data fusion strategy. Sequential belief mass fusion reconciles inconsistencies between content relevance and public reception. It provides a nuanced view of topic salience under uncertainty. Beyond its current application, the framework is remarkably adaptable. It could be extended to domains such as sentiment analysis, trend detection or scientific text mining. It also could benefit from further refinements, particularly in temporal dynamics and fusion rule optimization. Better conflict management within belief fusion, for example, would improve sensitivity to refine patterns and correlated topics. Overall, this approach provides a flexible, transferable and theoretically practical tool for quantitative analysis of digital discourse. It extensively relevant across fields that work with massive textual data and online influence.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Fatima-Zahrae Sifi	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Wafae Sabbar	✓	✓		✓	✓	✓	✓			✓		✓		
Amal El Mzabi	✓	✓		✓	✓	✓	✓			✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest regarding the publication of this article. The authors confirmed that the paper was free of plagiarism.

INFORMED CONSENT

This study does not involve human participants or personal data. Therefore, informed consent is not required in this study.

ETHICAL APPROVAL

This research does not involve human subjects or animals. Therefore, ethical approval was not required.

DATA AVAILABILITY

The authors confirm that the data supporting the findings of this study are available within the article [and/or its supplementary materials].

REFERENCES

[1] K. K. Kapoor, K. Tamilmani, N. P. Rana, P. Patil, Y. K. Dwivedi, and S. Nerur, "Advances in social media research: past, present and future," *Information Systems Frontiers*, vol. 20, no. 3, pp. 531–558, Jun. 2018, doi: 10.1007/s10796-017-9810-y.

[2] V. Barger, J. W. Peltier, and D. E. Schultz, "Social media and consumer engagement: a review and research agenda," *Journal of Research in Interactive Marketing*, vol. 10, no. 4, pp. 268–287, Oct. 2016, doi: 10.1108/JRIM-06-2016-0065.

[3] S. Zhou, M. Blazquez, H. McCormick, and L. Barnes, "How social media influencers' narrative strategies benefit cultivating influencer marketing: tackling issues of cultural barriers, commercialised content, and sponsorship disclosure," *Journal of Business Research*, vol. 134, pp. 122–142, Sep. 2021, doi: 10.1016/j.jbusres.2021.05.011.

[4] M. El Marrakchi, H. Bensaid, and M. Bellafkih, "E-reputation prediction model in online social networks," *International Journal of Intelligent Systems and Applications*, vol. 9, no. 11, pp. 17–25, Nov. 2017, doi: 10.5815/ijisa.2017.11.03.

[5] Y. Yan, F. Toriumi, and T. Sugawara, "Understanding how retweets influence the behaviors of social networking service users via agent-based simulation," *Computational Social Networks*, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40649-021-00099-8.

[6] B. Wilder, N. Immorlica, E. Rice, and M. Tambe, "Maximizing influence in an unknown social network," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018, doi: 10.1609/aaai.v32i1.11585.

[7] J. Tang and H. Liu, "Unsupervised feature selection for linked social media data," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, Aug. 2012, pp. 904–912, doi: 10.1145/2339530.2339673.

[8] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi, "Measuring user influence in Twitter: the million follower fallacy," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 4, no. 1, pp. 10–17, May 2010, doi: 10.1609/icwsm.v4i1.14033.

[9] S. Jendoubi and A. Martin, "A reliability-based approach for influence maximization using the evidence theory," in *Big Data Analytics and Knowledge Discovery (DaWaK 2017)*, 2017, pp. 313–326, doi: 10.1007/978-3-319-64283-3_23.

[10] H. Jelodar *et al.*, "Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15169–15211, Jun. 2019, doi: 10.1007/s11042-018-6894-4.

[11] T. Denœux, "Decision-making with belief functions: a review," *International Journal of Approximate Reasoning*, vol. 109, pp. 87–110, Jun. 2019, doi: 10.1016/j.ijar.2019.03.009.

[12] Y. Huang, Q. Liu, J. Liu, and Y. Hu, "Topic discovery in scientific literature," in *Computer Supported Cooperative Work and Social Computing (ChineseCSCW 2022)*, 2023, pp. 481–491, doi: 10.1007/978-981-99-2356-4_38.

[13] R. Jiroušek, V. Kratochvíl, and P. P. Shenoy, "Computing the decomposable entropy of belief-function graphical models," *International Journal of Approximate Reasoning*, vol. 161, Oct. 2023, doi: 10.1016/j.ijar.2023.108984.

[14] E. Koksalmis and Ö. Kabak, "Sensor fusion based on Dempster-Shafer theory of evidence using a large scale group decision making approach," *International Journal of Intelligent Systems*, vol. 35, no. 7, pp. 1126–1162, Jul. 2020, doi: 10.1002/int.22237.

[15] J. Ni, J. Luo, and W. Liu, "3D palmprint recognition using Dempster-Shafer fusion theory," *Journal of Sensors*, vol. 2015, pp. 1–7, 2015, doi: 10.1155/2015/252086.

[16] N. J. J. Smith, "Acting on belief functions," *Theory and Decision*, vol. 95, no. 4, pp. 575–621, Nov. 2023, doi: 10.1007/s11238-023-09937-9.

[17] Z. A. Sosnowski and J. S. Walijewski, "Fuzzy Dempster-Shafer modelling and decision rules," in *Computer Information Systems and Industrial Management: 15th IFIP TC8 International Conference*, 2016, pp. 516–529, doi: 10.1007/978-3-319-45378-1_46.

[18] R. Jiroušek, V. Kratochvíl, and P. P. Shenoy, "On conditional belief functions in directed graphical models in the Dempster-Shafer theory," *International Journal of Approximate Reasoning*, vol. 160, Sep. 2023, doi: 10.1016/j.ijar.2023.108976.

[19] P. P. Shenoy, "On distinct belief functions in the Dempster-Shafer theory," in *Proceedings of the Thirteenth International Symposium on Imprecise Probability: Theories and Applications*, 2023, pp. 426–437.

[20] Z. Liu, Y. Cao, X. Yang, and L. Liu, "A new uncertainty measure via belief Rényi entropy in Dempster-Shafer theory and its application to decision making," *Communications in Statistics - Theory and Methods*, vol. 53, no. 19, pp. 6852–6868, Oct. 2024, doi: 10.1080/03610926.2023.2253342.

[21] K. Qu, G. Zhao, Y. Wu, and L. Tong, "Research on airspace conflict detection method based on spherical discrete grid representation," *Applied Sciences*, vol. 13, no. 11, May 2023, doi: 10.3390/app13116493.

[22] D. P. Prieto, R. de Haan, and A. Özgün, "A belief model for conflicting and uncertain evidence: connecting Dempster-Shafer theory and the topology of evidence," in *Proceedings of the Twentieth International Conference on Principles of Knowledge Representation and Reasoning*, Sep. 2023, pp. 552–561, doi: 10.24963/kr.2023/54.

[23] P. Smets and R. Kennes, "The transferable belief model," *Artificial Intelligence*, vol. 66, no. 2, pp. 191–234, Apr. 1994, doi: 10.1016/0004-3702(94)90026-4.

[24] K. Sentz and S. Ferson, "Combination of evidence in Dempster-Shafer theory," *Reports*, Albuquerque, NM, and Livermore, CA (United States), Apr. 2002, doi: 10.2172/800792.

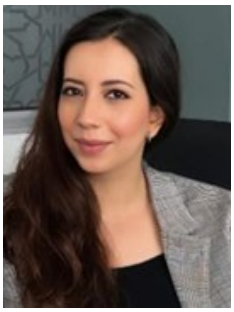
[25] B. Suo, L. Zhao, and Y. Yan, "A novel Dempster-Shafer theory-based approach with weighted average for failure mode and effects analysis under uncertainty," *Journal of Loss Prevention in the Process Industries*, vol. 65, May 2020, doi: 10.1016/j.jlp.2020.104145.

[26] P. Kumari, S. R. N. Reddy, and R. Yadav, "Application of Dempster-Shafer theory in sensor data fusion," in *Proceedings of the International Conference on Machine Learning, Deep Learning and Computational Intelligence for Wireless Communication*, 2024, pp. 91–102, doi: 10.1007/978-3-031-47942-7_9.

[27] L. Qiyanhui, "A framework for multi-modal fusion using Dempster-Shafer theory in computer vision applications," in *2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS)*, Jul. 2024, pp. 239–245, doi: 10.1109/ICPICS62053.2024.10797105.

[28] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
[29] D. Effrosynidis, S. Symeonidis, and A. Arampatzis, "A comparison of pre-processing techniques for Twitter sentiment analysis," in *Research and Advanced Technology for Digital Libraries*, 2017, pp. 394–406, doi: 10.1007/978-3-319-67008-9_31.
[30] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp. 77–84, Apr. 2012, doi: 10.1145/2133806.2133826.
[31] S. S. Kamaruddin, S. A.-Rahman, and W. Wibowo, "Understanding Malaysian public opinion on suicide through sentiment analysis and topic modeling of Reddit posts," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 18055–18062, Dec. 2024, doi: 10.48084/etasr.8738.
[32] M. M. Marchetti, "Tweets dataset," *Kaggle.com*. Accessed: Apr. 15, 2025. [Online]. Available: https://www.kaggle.com/datasets/mmmarchetti/tweets-dataset/data

BIOGRAPHIES OF AUTHORS



Fatima-Zahrae Sifi is a knowledge and data science engineer from the School of Information Sciences (ESI), Rabat, Morocco, in 2016. She also completed higher school preparatory classes specializing in Mathematics in Meknes, Morocco, in 2013. She has published several research papers in international conferences and journals, including E3S Web of Conferences (2021) and the 14th International Conference on Intelligent Systems: Theories and Applications (SITA) (2023). Her research interests include text mining, topic modeling, machine learning, deep learning, and graph theory. She can be contacted at email: fatimazahraesifi@gmail.com or fatimazahrae.sifi@univh2c.ma.

Wafae Sabbar received the Ph.D. degree from the Faculty of Science and Technology, Hassan II University, Mohammedia, Morocco, in 2006. She is currently a professor of Higher Education at the Faculty of Law, Economic and Social Sciences of Ain Sebaa (FSJESAS), Morocco. Her research interests include information technologies, computer science and AI. She can be contacted at email: swafae@gmail.com.

Amal El Mzabi received the Ph.D. degree from the Faculty of Science and Technology, Hassan II University, Mohammedia, Morocco, in 2005. She is currently a professor of Higher Education at the Faculty of Law, Economic and Social Sciences of Mohammedia (FSJESM), Morocco. Her research interests include information technologies, computer science and AI. She can be contacted at email: amal.elmzabi@gmail.com.