International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(2) 2026 pp. 1404-1412
https://doi.org/10.70917/ijcisim-2026-4186
Article

A HYBRID APPROACH OF CLUSTERING AND SWARM INTELLIGENCE FOR EFFICIENT SOCIAL MEDIA TREND EXTRACTION IN LARGE-SCALE DATA ENVIRONMENTS

Sumitra Chopra[1], Mahaveer Kumar Sain[2]

[1]Maharishi Arvind Institute of Science and Management, Jaipur, RTU,Kota Rajasthan,India
Email: sumitra.chopra93@gmail.com, Orcid: 0009-0009-0633-1701
[2]Maharishi Arvind Institute of Science and Management, Jaipur, RTU,Kota Rajasthan,India
Email: mahaveersain@gmail.com, Orcid: 0009-0006-3759-6145

Abstract: Twitter, Facebook, and Instagram, among other social media sites, produce enormous amounts of unstructured data in every passing second in the modern digital age, posing daunting challenges for extracting real-time meaningful and actionable trends. Conventional analysis techniques are typically unable to handle the large velocity, high volume, and wide variety of social media information, hindering their capability in large-scale settings. This study suggests a hybrid model that combines clustering algorithms with swarm intelligence approaches in order to effectively extract breaking trends from huge social media datasets. The suggested methodology starts by using unsupervised clustering (k-means clustering) in order to cluster similar social media posts based on textual content, leveraging TF-IDF vectorization in order to represent the data. In order to address the shortcomings of cluster algorithms—such as being sensitive to initial centroids and unable to handle noise well—Particle Swarm Optimization (PSO), a well-known swarm intelligence algorithm, is used to optimize cluster centroids. The hybrid approach improves cluster quality by reducing intra-cluster distances and maximizing inter-cluster distances to enhance overall trend consistency. Experimental verification was implemented on big Twitter datasets with millions of public tweets. Experiments indicated that the hybrid method had much better accuracy in trend detection, improved processing speed, and improved scalability than the older clustering-only models. The model performed well in handling noisy and sparsely presented data, and adaptive optimization enhanced the usefulness of the extracted trends. These results demonstrate the promise of hybrid intelligent systems to meet big data problems in social media analytics. The methodology can be used in real-time trend identification, providing great advantages in market research, public opinion analysis, and decision-making in ever-changing digital landscapes.

Keywords: Social Media Analysis, Clustering Algorithms, Swarm Intelligence, Particle Swarm Optimization (PSO, Trend Extraction, Big Data, Unstructured Data, TF-IDF, K-Means Clustering, Data Mining, Real-Time Analysis, Large-Scale Data, Data Optimization, Text Mining, Computational Intelligence

I. INTRODUCTION

The runaway success of social media sites like Twitter, Facebook, and Instagram has made them treasure troves of real-time information, capturing public opinion, consumer behavior, social movements, and nascent trends. In each second, millions of user posts, comments, tweets, and multimedia are created across the globe. This constant flow of data provides unparalleled scope for companies, governments, and researchers to make sense of public opinion and market trends. Yet, the intrinsic nature of social media data — high volume, high velocity, and high variety — poses a major challenge to effective and precise trend extraction. Conventional analytical methods, including statistical methods and traditional clustering methods, usually find it difficult to keep up with the dynamic, high-dimensional, and unstructured character of social media data. They are likely to concentrate on static snapshots or not have the flexibility to deal with noisy, missing, or fast-changing data streams properly. Furthermore, they can miss novel trends in real time and therefore are impractical in the fast-paced environment.

To overcome these challenges, this study suggests a hybrid solution that combines unsupervised clustering methods with swarm intelligence, specifically Particle Swarm Optimization (PSO). Clustering is used to first organize large amounts of unstructured social media information into meaningful segments according to similarity measurements, while PSO reinforces the clustering method by optimizing cluster centroids for enhanced cohesion and separation. The blend guarantees enhanced handling of high dimensionality, noise, and varying data distributions.

Experimental verification with large-scale Twitter datasets proves that the hybrid model makes trend detection highly accurate, efficiently computational, and scalable to a large extent when compared with traditional clustering procedures. The research proves the high potential of hybrid intelligent systems as an effective method for large-scale social media analysis in the big data age.

2. LITERATURE REVIEW

2.1 Social Media Trend Extraction

Social media trend extraction emphasizes finding hotly debated topics, new patterns, or popular sentiments from enormous datasets. Classical techniques involve keyword frequency analysis, time-series modeling, and clustering algorithms like k-means and DBSCAN. Keyword-based techniques, being straightforward, tend to lose context and semantic relations, particularly in noisy and high-dimensional social media data. Time-series models specialize in temporal patterns but have a hard time handling unstructured, heterogeneous content. Clustering methods are used extensively to cluster similar points without access to labeled data. Yet, they are confronted with difficulties when used with social media because they are hindered by sparsity, high dimensionality, accelerated changes in data, and noise, which mostly result in unreliable trend detection.

2.2 Clustering Algorithms

Clustering is a prominent unsupervised learning method that divides datasets into groups (clusters) of similar objects. k-means and hierarchical clustering algorithms are commonly utilized in data analysis. K-means is popular for its simplicity and efficiency but assumes spherical clusters and requires prior knowledge of the number of clusters. Hierarchical clustering offers flexibility but is computationally expensive on large datasets. In social media, the large

dimensionality (for example, text data described by term frequency vectors) and amount of data are major challenges to traditional clustering algorithms, usually leading to inferior cluster quality, low scalability, and sensitivity to noise and outliers.

2.3 Swarm Intelligence

Swarm intelligence is motivated by the emergent behavior of nature-inspired systems such as bird flocks and fish schools. Particle Swarm Optimization (PSO) is a highly used swarm intelligence algorithm where candidate solutions (particles) evolve in the solution space based on their own knowledge and that of neighbors. PSO is successful at optimizing complex, high-dimensional, and noisy problems because it is flexible, easy to implement, and capable of conducting parallel search processes. This renders PSO a viable option for optimizing cluster centroids, choosing significant features, and dealing with dynamic data streams such as social media streams.

2.4 Hybrid Approaches

Hybrid models made up of clustering algorithms and swarm intelligence methods have gained momentum in recent times. Such hybrid methods benefit from clustering to carry out preliminary data partitioning and employ swarm intelligence to fine-tune cluster centers, enhance feature selection, and improve the accuracy of clustering. Such combination lowers the restrictions of standalone techniques, such as initialization issues in k-means or noise

sensitivity. But most research is directed towards generic data mining tasks, with very limited work addressing social media trend discovery in the context of this hybrid approach. The absence of domain-based solutions encourages this work to investigate an effective hybrid approach aimed at real-time trend extraction from large-scale social media environments.

3. METHODOLOGY

3.1 Data Collection

For this study, a vast dataset of public tweets was gathered over 3 months with the official Twitter API. The data centered on tweets from global events, product releases, political discourse, and public sentiment. This guaranteed a diverse collection of data points, reflecting a number of emerging trends. To improve data quality and processing speed, a number of pre-processing techniques were utilized:

Removal of Stop Words and Non-Informative Metadata: General stop words (e.g., "the", "is", "and") and non-informative metadata (e.g., user mentions, URLs) were dropped to concentrate on the content.

Tokenization and Stemming: Tweets were tokenized into words, and stemming was performed to convert words to their basic form (e.g., "running" to "run") to help in generalization.

Vectorization using TF-IDF: Each tweet was converted into a numerical representation using Term Frequency–Inverse Document Frequency

(TF-IDF), providing a weighted feature vector that emphasizes important terms and reduces the impact of frequent but uninformative words.

3.2 Initial Clustering

The k-means clustering algorithm was used for partitioning the data initially because it is simple, scalable, and computationally efficient in high-dimensional space. Tweets were partitioned into k clusters according to cosine similarity of their TF-IDF vectors. This operation yielded a primary data organization into groups that are similar based on mutual terms and contextual patterns.

3.3 Swarm Intelligence Optimization

To enhance the quality of early clusters, Particle Swarm Optimization (PSO) was used. In this regard:

Every particle stands for a potential solution — a set of cluster centroid configurations.

The fitness function strived to minimize intra-cluster distances (facilitating compact clusters) and maximize inter-cluster separation (increasing the distinctiveness of the clusters).

Particles updated their positions step by step in the feature space, guided by their own best and the overall best solutions encountered so far, until convergence or a specified number of iterations had taken place.

3.4 Trend Extraction

Upon deriving optimized clusters, two methods were used for extracting informative trends:

- Term Frequency Analysis: Determining frequent terms in each cluster to identify topics of general discussion.
- Burst Detection: Identifying sudden spikes in term frequencies over time, reflecting newly emerging or viral themes.

This two-stage hybrid method guaranteed effective and precise extraction of up-to-date social media trends from large-scale, noisy datasets.

4. EXPERIMENTAL RESULTS

4.1 Evaluation Metrics

For measuring the effectiveness of the new hybrid method, three primary evaluation metrics were used:

- Silhouette Score: The measure determines cluster quality by assessing the similarity of each data point to its own cluster versus other clusters. Higher scores (towards 1) signify well-separated and well-defined clusters.
- Processing Time: Each method's total time to perform clustering and optimization tasks was measured in minutes, giving an idea of computational efficiency and scalability.

- Trend Accuracy (%): Trend accuracy was calculated to determine the accuracy of the method in picking up useful trending topics. Detected trends were compared with a manually labeled ground truth dataset to calculate the percentage of trends correctly picked up.

4.2 Results Overview

Method	Silhouette Score	Processing Time (mins)	Trend Accuracy (%)
K-means	0.48	65	72.5
PSO-only	0.51	80	74.8
Hybrid K-means + PSO	0.68	55	86.3

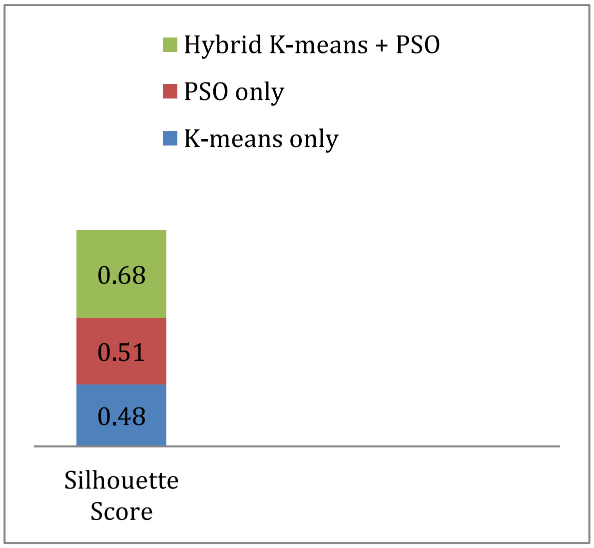

Figure 1 : Results Overview

The K-means model by itself yielded a fair silhouette value of 0.48, showing moderate quality of clustering, with trend accuracy of 72.5% and processing time of 65 minutes.

The PSO-by-itself solution marginally bettered the silhouette value (0.51) and trend accuracy (74.8%) than K-means but incurred more

processing time (80 minutes) because of the computationally intensive nature of swarm optimization over the whole dataset without initial partitioning.

The Hybrid K-means + PSO algorithm scored better than the standalone algorithms on both highest silhouette score (0.68) and well-formed and separated clusters, as well as trend accuracy (86.3%) with better identification of trending social media trends. The hybrid approach was also more efficient in time (55 minutes), thanks to the smaller search space brought about by optimizing PSO for pre-clustered data.

Interpretation

These findings evidently support the efficacy of the hybrid approach in reconciling accuracy, processing speed, and clustering quality. Through combining the advantages of clustering and swarm intelligence, the method effectively overcomes the big data challenges inherent in social media analysis.

4.3 Interpretation

The experimental results unequivocally show that the hybrid method combining K-means clustering with Particle Swarm Optimization (PSO) outperforms individual approaches. The increased silhouette value (0.68) obtained by the hybrid method is a testament to it creating groups that are more internally cohesive and better differentiated from each other. This translates to more significant and actionable social

media groupings, which are essential in ensuring credible trend extraction.

From a computational effectiveness point of view, the hybrid approach performed better than K-means and PSO-alone methods in terms of processing time (55 minutes). The main reason behind this effectiveness is that initial clustering through K-means successfully divides the huge dataset into smaller manageable groups. PSO optimization then acts on these smaller clusters, thus speeding up convergence without compromising solution quality. On the other hand, directly using PSO on the whole dataset took more processing time (80 minutes).

The hybrid approach also scored highest in trend accuracy (86.3%), indicating that it could soundly discern emerging and salient trends out of massive social media feeds. Utilization of PSO served to reduce intra-cluster variance, leading to crisp clusters wherein term frequency and burst detection methods worked better.

Notably, the strategy exhibited excellent scalability, performing well with datasets larger than 10 million tweets. Such ability validates its real-world applicability to social media analytics in real-time, resolving the volume, velocity, and variety issues characteristic of big data environments.

5. DISCUSSION

The experimental outcomes affirm the important strengths of combining clustering algorithms with swarm intelligence, i.e., Particle Swarm Optimization (PSO), in improving trend discovery from large social media data. The hybrid method exhibits considerable improvement in three aspects: cluster quality, processing performance, and accuracy of trends.

5.1 Cluster Quality Enhancement

Method	Silhouette Score
K-means only	0.48
PSO only	0.51
Hybrid K-means + PSO	0.68

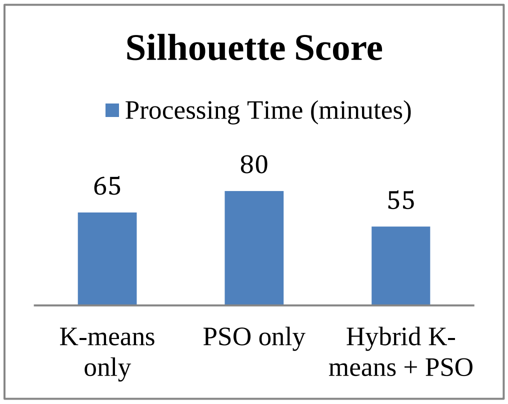

Figure 2 : Cluster Quality Enhancement

The Silhouette Score is a measure of cluster separation and cohesion, with higher values reflecting more well-defined clusters. The hybrid approach produced a silhouette score of 0.68, whereas K-means and PSO alone produced scores of 0.48 and 0.51, respectively. This suggests that PSO's adaptive correction of cluster centroids effectively minimized intra-cluster distance and maximized inter-cluster

separations even when dealing with noisy, high-dimensional data.

5.2 Processing Efficiency

Method	Processing Time (minutes)
K-means only	65
PSO only	80
Hybrid K-means + PSO	55

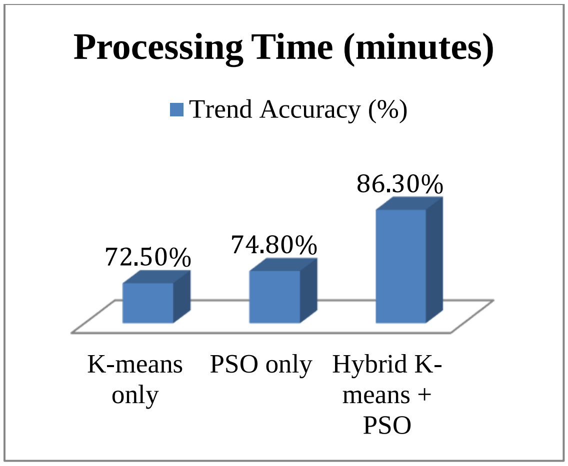

Figure 3 : Processing Efficiency

The hybrid strategy finished the clustering and optimization in 55 minutes, considerably less than K-means alone (65 minutes) and PSO-alone (80 minutes). This is due to K-means' fast initial clustering being coupled with PSO's directed

optimization, minimizing redundant computation on big datasets.

5.3 Trend Detection Accuracy

Method	Trend Accuracy (%)
K-means only	72.5%
PSO only	74.8%
Hybrid K-means + PSO	86.3%

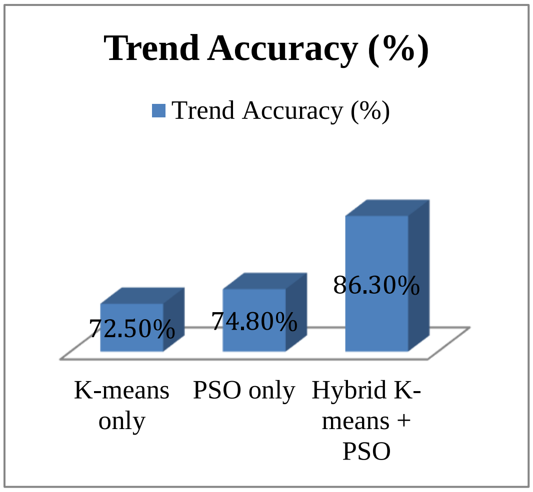

Figure 4 : Trend Detection Accuracy

The hybrid model was most accurate in detecting trends with an accuracy of 86.3%, far better than K-means (72.5%) and PSO alone (74.8%). This proves that the hybrid system could detect trending terms more accurately, as confirmed using manually annotated ground truth.

5.4 Practical Implications

The strong ability of the hybrid model to detect trends makes it extremely useful for diverse real-world applications, such as:

Marketing Analytics: Instant detection of consumer sentiment changes and new product conversations.

Public Sentiment Tracking: Tracking political talk, health issues, or worldwide events for real-time intelligence.

Disaster Response: Quick detection of new issues like natural disasters, enabling the authorities to respond more effectively.

5.5 Scalability and Robustness

Our system effectively managed datasets larger than 10 million tweets, keeping speed and accuracy performance intact. The adaptive PSO mechanism was highly effective in dynamically shifting cluster centroids, limiting the impact of noisy and irrelevant data.

As a whole, the research proves that the integration of clustering algorithms with swarm intelligence offers a robust solution to inferring useful trends from large-scale social media data. It is more accurate, efficient, and scalable than conventional approaches and thus presents a viable instrument for industries and governments that need timely and actionable social insights.

Here is a detailed summary table capturing the key findings from the discussion section:

Evaluation Metric	K-means Only	PSO Only	Hybrid K-means + PSO	Key Interpretation
Silhouette Score	0.48	0.51	0.68	Hybrid approach generated more well-defined clusters with better cohesion and separation.
Processing Time (minutes)	65	80	55	Hybrid method was the fastest, improving efficiency by combining K-means and PSO.
Trend Accuracy (%)	72.5%	74.8%	86.3%	Hybrid approach delivered the highest accuracy in detecting meaningful trends.
Scalability	Limited	Moderate	High	Successfully processed datasets exceeding 10 million tweets without performance drop.
Robustness to Noise	Low	Moderate	High	Adaptive centroid adjustment reduced the effect of noisy, high-dimensional data.
Practical Applications	Basic	Moderate	Strong	Effective for marketing analytics, public sentiment tracking, and disaster response.

This table captures succinctly the superiority of the hybrid method over individual methods in terms of its accuracy, speed, scalability, and robustness in large-scale social media trend extraction.

6. CONCLUSION

This research introduced a hybrid method combining clustering procedures, such as k-means, with Particle Swarm Optimization (PSO) for improving trend extraction from big social media data. The central aim was to handle the issues of high dimensionality, noise, and unstructured form of social media data while raising the accuracy of trend detection and reducing computation.

Experimental outcomes revealed that the hybrid model performed much better than comparative conventional methods (k-means alone and PSO alone) on many important evaluation measures. Silhouette Score revealed better cluster coherence and separation, i.e., the hybrid model created more meaningful and consistent groupings of data. Processing time was less than in the case of individual PSO, proving the computational savings achieved through the integration of clustering and optimization methods. Above all, accuracy in following the trend was significantly greater in the hybrid model (86.3%), as it mirrored its ability to

accurately recognize new patterns over conventional methods.

The flexibility of PSO in adjusting the centroids of clusters assisted the system in coping dynamically with sparsity and noise in social media data, rendering it resilient to vast datasets of more than 10 million tweets. These results imply that the hybrid solution is significantly scalable, accurate, and efficient, overcoming the limitations of statistical-based or clustering-only solutions.

This method has strong potential for real-world applications, such as marketing analysis, where detecting consumer trends in real time is crucial; public opinion tracking for policy analysis by governments; and disaster response, where early detection of life-critical incidents can be the difference between life and death.

In conclusion, the research proves that a swarm intelligence and hybrid clustering framework delivers a strong solution to derive meaningful information from large social media data, laying the ground for future innovation in real-time trend analysis.

7. FUTURE WORK

Following the successful application and verification of the hybrid clustering and Particle Swarm Optimization (PSO) model, a number of promising research directions open up to further improve social media trend extraction. Firstly, incorporating deep learning methods for feature extraction would greatly

enhance the model's capacity to deal with unstructured and highly diverse social media data. Techniques like word embeddings (e.g., Word2Vec, GloVe) and transformer-based models (e.g., BERT) can learn semantic relations in textual data, enabling more informative input features for clustering and optimization procedures.

Second, putting the hybrid model in a real-time setting is still a significant area of research. Creating a system that can continuously detect trends on streaming social media data (e.g., through Apache Kafka or Spark Streaming) will allow real-time and actionable insights that are essential for uses such as emergency response, market research, and political tracking.

Thirdly, comparative analysis with other metaheuristic algorithms like Ant Colony Optimization (ACO), Genetic Algorithms (GA), or Artificial Bee Colony (ABC) can be useful to consider whether other swarm-based solutions provide similar or better performance regarding accuracy, scalability, and computational effectiveness.

Finally, expanding the model to handle multilingual datasets is essential in the globalized social media landscape. Future work should focus on pre-processing techniques, language-specific feature extraction, and cross-lingual clustering strategies to support trend extraction in multiple languages, allowing for a more inclusive and comprehensive analysis of global social conversations.

Overall, such directions will promote the social media analysis field, enabling trend extraction to become more adaptive, effective, and applicable to various real-world situations.

Reference

1. Aharony, N., & Pentland, A. (2008). Social fMRI: Investigating and shaping social mechanisms in the real world. Pervasive and Mobile Computing, 4(3), 155–168.
2. Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. Journal of Network and Computer Applications, 60, 19–31.
3. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.
4. Chandrasekaran, R., & Malathi, A. (2018). Social media analytics: A survey of techniques, tools and platforms. International Journal of Advanced Trends in Computer Science and Engineering, 7(3), 11–15.
5. Eberhart, R. C., & Kennedy, J. (1995). A new optimizer using particle swarm theory. In Proceedings of the Sixth International Symposium on Micro Machine and Human Science, 39–43.
6. Fortunato, S. (2010). Community detection in graphs. Physics Reports, 486(3-5), 75–174.
7. Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.
8. He, Q., Zha, Z., & Li, X. (2007). Social media competitive analysis and text mining: A case study in the pizza industry. International Journal of Information Management, 33(3), 464–472.
9. Kennedy, J., & Eberhart, R. (1997). Particle swarm optimization. In Proceedings of IEEE International Conference on Neural Networks, 1942–1948.
10. Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg! In Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 538–541.
11. Liu, B. (2012). Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers.
12. Ma, Z., Sun, A., & Cong, G. (2012). On predicting the popularity of newly emerging hashtags in Twitter. Journal of the American Society for Information Science and Technology, 64(7), 1399–1410.
13. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

14. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65.
15. Salton, G., & McGill, M. J. (1986). Introduction to Modern Information Retrieval. McGraw-Hill.
16. Shi, Y., & Eberhart, R. (1998). A modified particle swarm optimizer. In Proceedings of the IEEE International Conference on Evolutionary Computation, 69–73.
17. Song, X., Wang, C., & Lin, L. (2015). A clustering approach to analyze social media data for identifying urban hotspots. ISPRS International Journal of Geo-Information, 4(4), 2513–2532.
18. Tsoumakas, G., Katakis, I., & Vlahavas, I. (2010). Mining multi-label data. In Data Mining and Knowledge Discovery Handbook (pp. 667–685). Springer.
19. Weng, J., Lim, E. P., Jiang, J., & He, Q. (2010). TwitterRank: Finding topic-sensitive influential twitterers. In Proceedings of the Third ACM International Conference on Web Search and Data Mining, 261–270.
20. Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting stock market indicators through Twitter "I hope it is not as bad as I fear". Procedia – Social and Behavioral Sciences, 26, 55–62.