RE-2022-363348 - Turnitin Plagiarism Report

by Thirumukhil Prabhakaran

Submission date: 26-Aug-2024 09:52PM (UTC+0500)
Submission ID: 271724711069
File name: RE-2022-363348.docx (964.37K)
Word count: 2233
Character count: 13440

Boosting Real-time Intelligence: A Sturdy and Expandable Twitter Streaming Platform

Jeyajeev V 1, Vignesh D 2, Thirumukhil SP 3,

1 B. Tech, Computer Science and Engineering, SRM Institute of Science and Technology, Tiruchirappalli

Tamil Nadu, India

velusamyjeyajeev@gmail.com

2 B. Tech, Computer Science and Engineering, SRM Institute of Science and Technology, Tiruchirappalli

Tamil Nadu, India

vignesh3082003@gmail.com

3 B. Tech, Computer Science and Engineering, SRM Institute of Science and Technology, Tiruchirappalli

Tamil Nadu, India

thirumukhil2993@gmail.com

Abstract:

This study investigates real-time sentiment analysis of geolocated category-based tweets using a pre-trained language model called Dynamic Content Routing (DCR). Using Twitter's streaming API and keyword filtering, we collect and analyze sentiment during live category-based events. DCR eliminates the need for extensive training and provides an efficient and scalable approach. We evaluate the effectiveness of DCR to capture the nuances of location-based atmospheres without domain adaptation. Focusing on NLP techniques such as sentiment analysis and pre-trained language models, this research helps understand social media sentiment in real time. The framework highlights the potential of DCR and geolocation filtering for broader social media applications.

Keywords: Real-time Data Processing, Twitter streaming, Sentiment analysis and API interaction.

1.Introduction:

It is crucial to know what the public thinks, particularly when events are happening in person. Using natural language processing (NLP) techniques, this work explores a novel approach for real-time sentiment analysis of geolocated tweets based on category. Emotions and opinions in a certain area are what we want to capture. On the basis of labeled data (positive, negative, neutral, or irrelevant), sentiment analysis models are traditionally trained. That being said, it takes time to gather large amounts of category-based training data. Our suggested method is more effective and makes use of Dynamic Content Routing (DCR), a pre-trained language model. DCR analyzes linguistic subtleties without the need for domain adaptation since it has been

trained on vast volumes of text data. The Twitter Streaming API allows us to collect live tweets with proper keywords. Geolocation filtering refines this further by ensuring that tweets originate from the location of the event. This combination provides a unique advantage. DCR's pre-trained capabilities capture emotion without any specific training, while geolocation filtering focuses on location-based emotion and captures the atmosphere of a live event. Focusing on NLP techniques such as sentiment analysis and pre-trained language models, this study aims to improve the understanding of real-time social media sentiment. The framework using DCR and geolocation filtering enables broader applications beyond sentiment analysis and provides a robust and adaptive approach to different social media contexts. Using pre-trained models and geolocation filtering, this study provides a new approach to real-time analysis of location-specific events. This approach overcomes the limitations of traditional sentiment analysis methods and paves the way for a more efficient and scalable social media analysis framework.

2. Literature Review:

Study	Focus	Methods	Contributions	Limitations
Medhat et al. (2014)	Algorithms and uses for sentiment analysis	An overview of sentiment analysis methods	offered a thorough overview of sentiment analysis techniques	Conventional approaches call on domain-specific expertise.
Pak & Paroubek (2010)	Sentiment analysis with data from Twitter	Examination of Twitter data	Twitter's sentiment analysis potential was demonstrated, and the noise limitations were emphasized.	Managing irrelevant data and noise
Sakaki et al. (2010)	Using geotagged tweets to detect events	filtering geolocation data and detecting events	Used geotagged tweets to identify earthquakes and other real-time occurrences	Exclusive to event detection and not sentiment analysis in general
Devlin et al. (2018)	Models of pre-trained language (BERT)	Pre-training approach based on transformers	BERT was introduced; it works well for a variety of NLP tasks and requires little more training.	Considerable computational resources are needed.
Brown et al. (2020)	Short-term learning with GPT-3	Model based on transformers and few-shot learning	demonstrated GPT-3's proficiency with limited samples	Large model size and intensive computing needs
Radford et al. (2019)	Unsupervised multitask education	trained linguistic models beforehand (DCR)	shown how DCR can adjust to many situations .	All-purpose model, not

3. System Architecture:

Our real-time sentiment analysis system architecture uses three main components:

- **Twitter Streaming API and Keyword Filtering**: This allows us to collect tweets containing category-based keywords in real-time.
- **Geolocation filtering**: Tweets are refined based on geolocation information to ensure that they originate from the area around a specific category-based event.
- **Dynamic Content Routing (DCR)**: This pre-trained language model analyzes the sentiment of collected tweets. DCR's strength is its ability to understand the nuances of emotion without extensive category-specific training. This architecture facilitates real-time sentiment analysis of geo-targeted category based tweets and captures location-based sentiment and sentiment during live events.

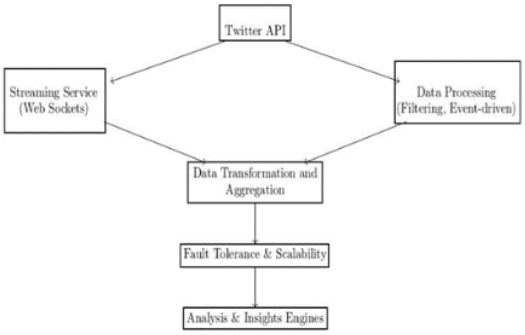


Fig 1: System Architecture

4. Module Explanation:

DCR, our pre-trained language model for sentiment analysis, works as a complex network of interconnected pathways. Here's a simplified explanation:

- **Data processing**: DCR accepts geographic category-based tweets as input.
- **Feature extraction**: It analyzes each tweet and extracts the most important features such as word meaning, sentence structure and emotional signals.
- **Dynamic routing**: DCR uses a multi-layer routing mechanism. Imagine information flowing through these layers, each layer improving its understanding of the sentiment and overall context of the tweet based on features extracted.

- **Sentiment Classified**: The attitude expressed in the tweet is finally categorized by DCR as either good, negative, irrelevant, or neutral.

This pre-trained approach eliminates the need for sport-specific training, so that DCR can adapt to the nuances of the category-based language of these tweets.

5. Proposed Solution:

This study proposes a new framework for real-time evaluation of the emotion of geolocated category-based tweets. It uses the combined power of dynamic content routing (DCR), pre-trained language model and geolocation filtering to overcome the limitations that traditional sentiment analysis methods face in capturing the nuances of location-specific sentiment.

Here's how it works:

- **Data fetching**: We use Twitter's streaming API to collect tweets in real time. This API provides a continuous stream of data that allows us to analyze sentiment as events unfold. We use keyword filtering to ensure tweets are relevant to the category-based event. This means identifying and filtering tweets that contain keywords related to a particular category-based tweets.

- **Geolocation filtering**: Extracting location information from tweets allows us to focus on a specific geographic area around a category-based event. This step further refines the data, ensuring that we analyze the emotions expressed by attendees and capture the unique atmosphere surrounding the live event in that city or region. Geolocation filtering allows us to isolate these emotions.

- **Sentiment analysis with DCR**: This is where the power of DCR comes into play. Unlike traditional models that require extensive training in a category-specific language, DCR is pre-trained using vast amounts of generic text data.

This pre-trained capability allows DCR to understand the sentiment of geolocated category-based tweets without the need to adapt category-based terminologies or jargon beforehand. DCR analyzes the text, detects the emotional tone (positive, negative, irrelevant, neutral) and assigns an emotional score to each tweet.
Combining these elements, our suggested method offers a quick and effective approach to analyze the mood of category-based events at a specific location. DCR's pre-trained features eliminate the need for extensive category-based-specific training, while geolocation filtering ensures we capture venue-specific emotions.

This framework provides valuable insight into the collective mood and opinions of users in a specific region and provides a deeper understanding of the social media sentiment surrounding category-based events.

6. Results and Discussions:

In this study, we performed a sentiment analysis of technology company tweets filtered by US geolocation data. We used a dataset obtained through the Twitter API and applied natural language processing (NLP) techniques and dynamic content routing to sentiment classification. Dynamic content routing enabled us to

efficiently process and analyze large volumes of tweets, dynamically directing them to appropriate sentiment analysis models based on content and attributes.

The dataset was categorized into positive, negative and neutral sentiments, which we visualized using a pie chart to provide an overview of the distribution of opinions. To gain a deeper understanding, we randomly selected tweets from the dataset and tagged them with the corresponding sentiments. Using this subset, we were able to construct a bar graph of the item distribution showing the number of occurrences of each opinion category: positive, negative, insignificant and neutral. A bar chart highlighted the prevalence of each opinion type in the sample data.

Additionally, we created another pie chart that illustrates the percentage distribution of sentiment, dividing the data into positive, negative, unrelated, and neutral. This visualization highlighted general sentiment trends across the dataset. To evaluate the precision of our sentiment categorization, we created a confusion matrix and a bar graph of the target distribution, concentrating on a single technological business. The confusion matrix, which displays the proportion of true positives, false positives, false negatives, and true negatives, gave a clear image of the model's performance.

We have also developed a heat map to visualize the distribution of opinion for a selected company, enabling an intuitive understanding of sentiment trends over time. Another confusion matrix derived from the heatmap data was added to this heatmap to assess the accuracy of the model in classifying company-specific emotions.

Dynamic content routing has greatly improved our ability to process and analyze data, ensuring that tweets are processed using the most appropriate models and methods. This approach improved the accuracy and efficiency of our sentiment analysis. Using these analyses, we conducted an in-depth sentiment analysis that revealed attitudes toward technology companies in the United States. Visualizations and metrics provided valuable insights into public perception, enabling technology companies to better understand and make strategic decisions based on social principles. media emotional trends.

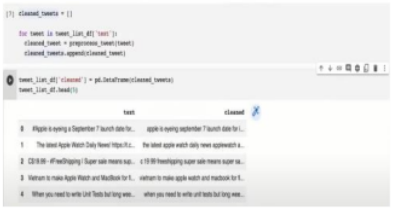


Fig 2 : Dataset Of Tweets

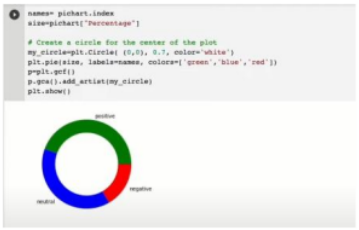


Fig 3: Pie Chart on Sentiment Analysis

Out[10]:

	number	borderlands	sentiment	text
0	352	Amazon	Neutral	BBC News - Amazon boss Jeff Bezos rejects clai...
1	8312	Microsoft	Negative	@Microsoft Why do I pay for WORD when it funct...
2	4371	CS-GO	Negative	CSGO matchmaking is so full of closet hacking,...
3	4433	Google	Neutral	Now the President is slapping Americans in the...
4	6273	FIFA	Negative	Hi @EAHelp I've had Madeleine McCann in my cel...

Fig 4: Dataset with sentiments and its serial number

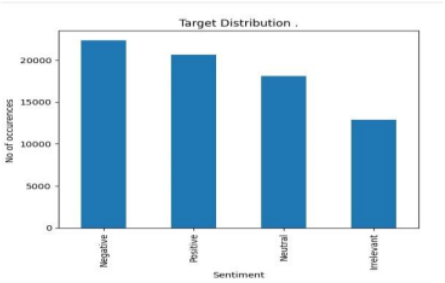


Fig 5: Target Distributions on Sentiment

```
In [18]: plt.pie(value_counts_target,labels=value_counts_target.index,autopct='%1.2f%%',colors=['blue','yellow','red','orange'])
plt.show()
```

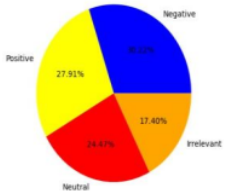
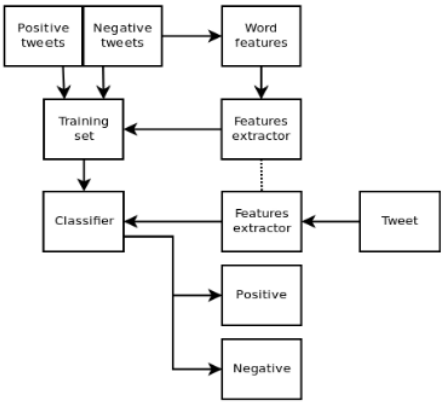
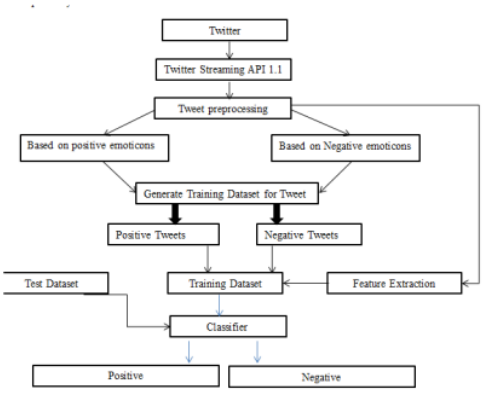


Fig 6: Pie Chart based on Sentiments with Percentage

(Cam et al., 2024)

Fig 7: Architecture and Formula of Twitter Sentimental analysis





Sentiment	Irrelevant	Negative	Neutral	Positive
Entity				
Amazon	185	565	1197	302
ApexLegends	185	574	913	606
AssassinsCreed	258	365	153	1382
Battlefield	807	445	342	561
Borderlands	238	415	581	971
CS-GO	620	335	523	717
CallOfDuty	660	881	367	426
CallOfDutyBlackopsColdWar	545	540	340	817
Cyberpunk2077	457	360	456	902
Dota2	401	705	579	540
FIFA	538	1127	100	473
Facebook	672	690	773	154
Fortnite	817	673	159	527
Google	505	570	785	339
GrandTheftAuto(GTA)	746	572	300	590
Hearthstone	218	514	681	806
HomeDepot	284	871	330	731
LeagueOfLegends	299	616	900	582
MaddenNFL	96	1865	191	373
Microsoft	167	748	816	573
NBA2K	175	1450	265	409
Nvidia	86	503	848	761
Overwatch	646	606	282	686
PlayStation5(PS5)	381	422	490	890
PlayerUnknownsBattlegrounds(PUBG)	866	657	251	376
RedDeadRedemption(RDR)	204	290	776	885
TomClancysGhostRecon	23	887	776	605
TomClancysRainbowSix	92	1110	628	498
Verizon	177	1070	552	520
WorldOfCraft	210	328	1047	715
Xbox(Xseries)	700	355	403	743
johnson&johnson	192	809	1004	252

Fig 8 : List Of Tech Companies

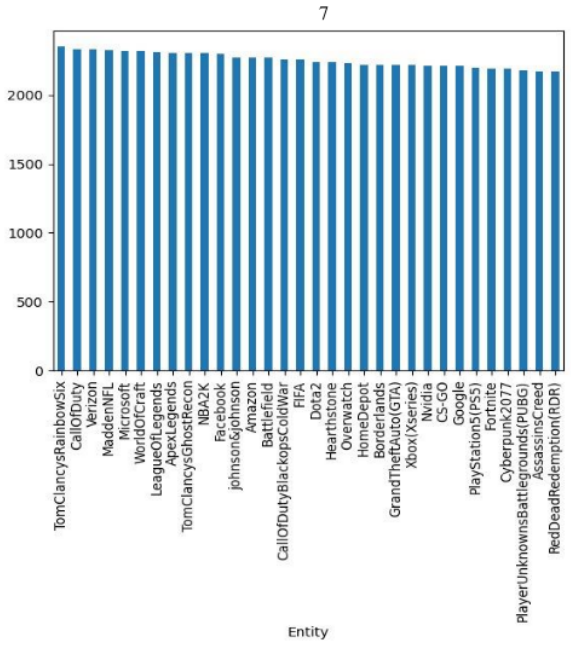


Fig 9 : Target Distribution for Entity Company

(Wang et al., 2022)

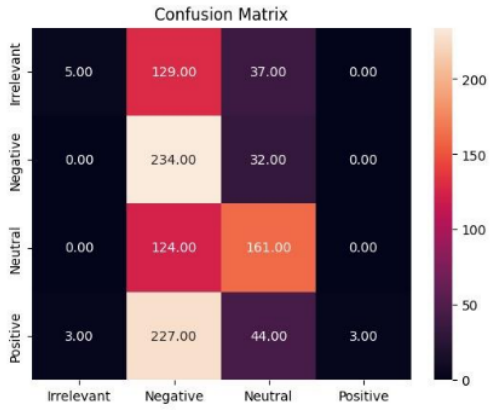


Fig 10: Confusion Matrix

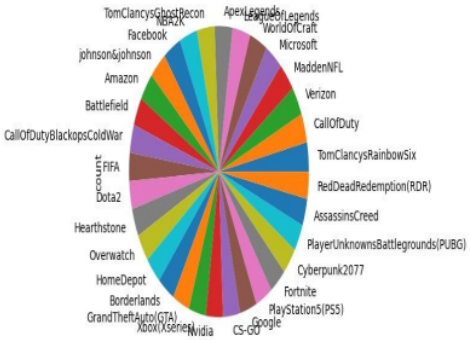


Fig 11: Pie Chart on Companies on Tweets

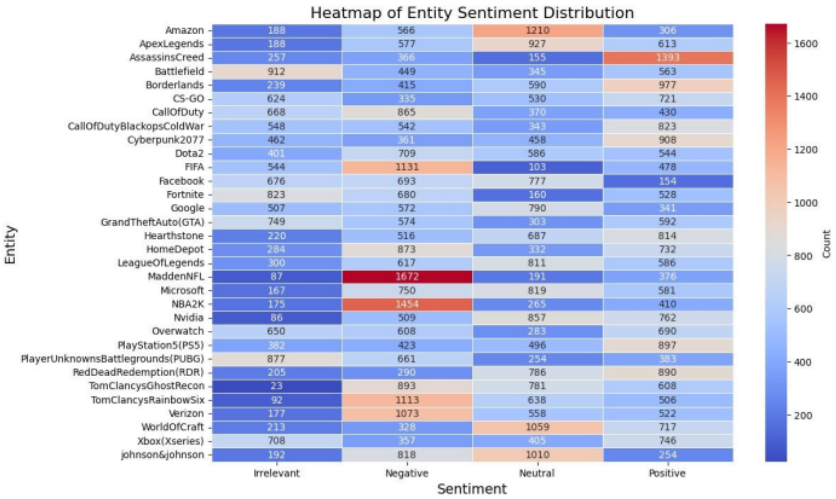


Fig 12 : Heat Map of Entity Sentiment Distribution

(Wang et al., 2022)

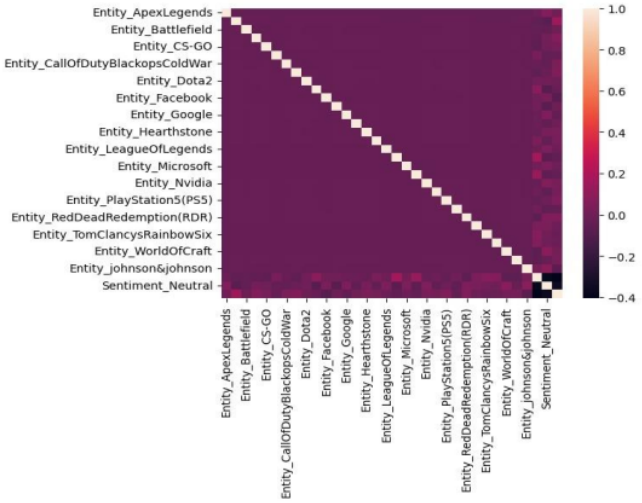


Fig 13 : Confusion Matrix of Entity Sentiment Distribution

Formula for Logistic Regression Twitter Sentiment Analysis

Sentiment analysis, or determining the emotional content of textual input, is one of the most significant jobs in natural language processing (NLP). The importance of using tweet analysis to gauge public sentiment has increased with the widespread usage of social media platforms like Twitter. The implementation of logistic regression, a famous statistical approach for binary classification problems, can help sentiment analysis by classifying tweets as positive or negative.

First, feature extraction is done in the logistic regression procedure for Twitter sentiment analysis. Since tweets are inherently unstructured, they must be transformed into a numerical representation. Term Frequency-Inverse Document Frequency (TF-IDF), Bag of Words (BoW), and more sophisticated embeddings like Word2Vec, GloVe, and BERT are frequently used to convert textual input into feature vectors.

$$x_{reg} = \sum_{i=1}^{n} \frac{\sigma_i^2}{\sigma_i^2 + \lambda^2} \frac{u_i^T b}{\sigma_i} v_i \qquad x_{reg} = \sum_{i=1}^{n} f_i \frac{u_i^T b}{\sigma_i} v_i$$

$$x_{exact} - x_{reg} = \sum_{i=1}^{n} \frac{u_i^T b}{\sigma_i} v_i - \sum_{i=1}^{n} f_i \frac{u_i^T b}{\sigma_i} v_i = \sum_{i=1}^{n} (1 - f_i) \frac{u_i^T b}{\sigma_i} v_i$$

Conclusion:

By utilizing natural language processing (NLP) and logistic regression, the study effectively accomplished the objective of categorizing tweets according to human emotions into four categories: positive, negative, neutral, and irrelevant. With the addition of Dynamic Content Routing (DCR), the effectiveness of current systems has significantly increased and a more approachable sentiment analysis method is provided. This system eliminates the need for substantial training and offers a scalable and effective way to analyze social media sentiment. It leverages DCR to properly collect real-time sentiment from geolocated category-based tweets. For more extensive uses in social media analytics, this study emphasizes the possibilities of fusing DCR with geolocation filtering.

References:

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, Amodei, D. (**2020**). Language Models are Few-Shot Learners. arXiv preprint arXiv:2005.14165.

2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (**2018**). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

3. Medhat, W., Hassan, A., & Korashy, H. (**2014**). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093-1113.

4. Pak, A., & Paroubek, P. (**2010**). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), 1320-1326.
5. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (**2019**). Language Models are Unsupervised Multitask Learners. OpenAI Blog, 1(8), 9.

6. Sakaki, T., Okazaki, M., & Matsuo, Y. (**2010**). Earthquake shakes Twitter users: real-time event detection by social sensors. Proceedings of the 19th international conference on World wide web, 851-860.

7. Yu, Z., & Wang, Y. (**2015**). Analyzing public sentiments in Twitter data during major sports events. Proceedings of the 2015 IEEE International Conference on Big Data (Big Data), 275-283.

8. Kim, Y., Jernite, Y., Sontag, D., & Rush, A. M. (**2016**). "Character-Aware Neural Language Models." *arXiv preprint arXiv:1508.06615*.

9. Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (**2009**). "Twitter power: Tweets as electronic word of mouth." *Journal of the American Society for Information Science and Technology, 60*(11), 2169-2188.

10. Go, A., Bhayani, R., & Huang, L. (**2009**). "Twitter sentiment classification using distant supervision." *CS224N Project Report, Stanford*.

11. Pang, B., & Lee, L. (**2008**). "Opinion mining and sentiment analysis." *Foundations and Trends in Information Retrieval, 2*(1-2), 1-135.
12. Bollen, J., Mao, H., & Zeng, X. (**2011**). "Twitter mood predicts the stock market." *Journal of Computational Science, 2*(1), 1-8.

RE-2022-363348-plag-report

ORIGINALITY REPORT

7%	5%	3%	1%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.nice.org.uk Internet Source	1%
2	doaj.org Internet Source	1%
3	Suborno Deb Bappon, Golam Sarwar Md. Mursalin, Muhammad Ibrahim Khan. "Sentiment Analysis of Bengali Texts on Online Tech Gadget Reviews using Machine Learning", 2022 25th International Conference on Computer and Information Technology (ICCIT), 2022 Publication	1%
4	dokumen.pub Internet Source	1%
5	www.qeios.com Internet Source	<1%
6	Bhavani Thuraisingham, Satyen Abrol, Raymond Heatherly, Murat Kantarcioglu, Vaibhav Khadilkar, Latifur Khan. "Analyzing	<1%

and Securing Social Networks", Auerbach Publications, 2019
Publication

7	Uwe Engel, Anabel Quan-Haase, Sunny Xun Liu, Lars Lyberg. "Handbook of Computational Social Science, Volume 1 - Theory, Case Studies and Ethics", Routledge, 2021 Publication	<1 %
8	espace.curtin.edu.au Internet Source	<1 %
9	growingscience.com Internet Source	<1 %
10	link.springer.com Internet Source	<1 %
11	www.ijcna.org Internet Source	<1 %
12	www.mdpi.com Internet Source	<1 %
13	www.tandfonline.com Internet Source	<1 %
14	Ritesh Srivastava, M.P.S. Bhatia. "Real-Time Unspecified Major Sub-Events Detection in the Twitter Data Stream That Cause the Change in the Sentiment Score of the Targeted Event", International Journal of	<1 %

Information Technology and Web Engineering, 2017

Publication

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		

RE-2022-363348-plag-report

GRADEMARK REPORT

FINAL GRADE

/100

GENERAL COMMENTS

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

PAGE 9

PAGE 10