Investigating the Locational Accuracy of Crowdsourced Data in the Context of NSW Wildfires Using NLP and Geocoding Techniques

Duminda Vishwajith
Department of Remote sensing and GIS
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
dumindavishwajith88@gmail.com

Saman Koswatte
Department of Remote sensing and GIS
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka
sam@geo.sab.ac.lk

***Abstract*- Wildfires are major natural disasters, and timely monitoring is essential for effective response. Recent progress in Crowdsourced Data (CSD) and Remote Sensing (RS) has improved wildfire monitoring, however their combined use for spatial validation is still limited. This study examined the locational accuracy of Twitter-based wildfire reports during the 2019-2020 New South Wales (NSW) - Australia wildfire season by comparing them with satellite-derived hotspots. Wildfire-related tweets were processed using Natural Language Processing (NLP), Named Entity Recognition (NER), and geocoding through the Nominatim API, while hotspot data was obtained from the VIIRS sensor via Digital Earth Australia. Advanced data cleaning and keyword filtering methods were applied to extract relevant geolocated tweets from a large raw dataset, and geocoded locations were rigorously clipped to the NSW boundary to improve spatial relevance. A 5 km grid overlay and statistical tests, including chi-square, binomial tests and Poisson-Based tests, were used to measure the spatial relationship between the two datasets. The results showed a strong positive association, indicating that Twitter reports often align with satellite-validated hotspots, particularly in densely reported areas. Although CSD can be uneven in coverage and limited by geocoding accuracy, a combined approach using NLP, spatial filtering, and statistical validation improves its reliability. The study highlighted the value of integrating social media data with RS, proposed a reproducible framework for spatial accuracy assessment, and provided practical guidance relevant to real-time wildfire monitoring and disaster management systems.**

Keywords- Crowdsourced Data (CSD), NER, NLP, Wildfire Monitoring, New South Wales (NSW)

I. INTRODUCTION

Wildfires are the most serious natural hazards worldwide and are becoming more frequent, intense, and prolonged due to climate change. Rising temperatures, extending dry periods, and frequent droughts create favorable conditions for large fires to spread. These events cause severe ecosystem damage, economic losses, and health impacts. Studies show that fire seasons have lengthened in recent decades, while smoke from major fires significantly reduces air quality and increases mortality, extending wildfire impacts far beyond burned areas [1].

Australia is one of the world's most fire-prone countries, where bushfires are a natural yet increasingly dangerous environmental process. Large areas burn annually because of flammable vegetation, extreme weather, and human ignition sources [2]. Major events such as Ash Wednesday (1983), Black Saturday (2009), and the 2019-2020 "Black Summer" demonstrate the severity of Australian fire seasons. New South Wales (NSW), with dense coastal populations and extensive bushland, was among the most affected regions and is important for evaluating advanced wildfire monitoring approaches [3], [4].

Satellite remote sensing plays a key role in wildfire monitoring by enabling detection of active fires and burned areas across large and inaccessible regions [5], [6]. Sensors such as MODIS and VIIRS provide near-real-time fire information, while Landsat and Sentinel support detailed burn assessments. In Australia, the Digital Earth Australia (DEA) Hotspots system uses MODIS and VIIRS data to support emergency management [7]. However, satellite observations are limited by revisit times, cloud and smoke cover, spatial resolution, and limited representation of local impacts on communities and infrastructure [8].

Alongside satellite systems, crowdsourced data provide new opportunities for collecting real-time ground-level wildfire information [9]. Social media platforms such as Twitter (now 'X') enable rapid sharing of fire sightings, smoke conditions, road closures, and evacuation updates. When processed effectively, these data can complement official datasets by filling spatial and temporal gaps and improving situational awareness [10]. During the 2019-2020 NSW wildfires, Twitter activity related to fire events increased sharply, highlighting its value for wildfire monitoring.

Despite this potential, the operational use of crowdsourced data remains challenging because of concerns about data quality and locational accuracy [11]. Social media content is often unstructured and may contain rumors, outdated information, or ambiguous place references. Previous studies also report uneven spatial coverage and difficulties integrating social media data with official spatial datasets [12].

Natural Language Processing (NLP), Named Entity Recognition (NER), and geocoding provide effective approaches to address these challenges. Transformer-based NER models such as BERT can identify location references in informal text, including abbreviations and local place names [13]. Gazetteer-based methods improve extraction by linking place names to geographic databases, while geocoding tools such as Nominatim convert these locations into coordinates for comparison with satellite-detected hotspots in a GIS environment. Statistical tests, including Chi-square and probability analyses, can then evaluate the alignment between crowdsourced and satellite fire detections [14].

https://doi.org/10.54389/GVDM5731

Although previous studies have explored the use of crowdsourced data and social media for disaster monitoring, limited research has focused on validating the locational accuracy of wildfire-related social media information using satellite observations. Most studies mainly examined information sharing, sentiment analysis, or general disaster communication rather than spatial reliability assessment. This study addresses that gap by integrating NLP, transformer-based NER, geocoding, and statistical validation within a unified framework to compare Twitter-derived locations with VIIRS wildfire hotspots during the 2019-2020 NSW Black Summer fires.

II. Methods

Wildfire hotspot data from the VIIRS sensor (via DEA) and Twitter data from September-December 2019 were processed for spatial comparison (Fig. 1). Twitter text was cleaned and filtered using wildfire-related keywords. Geographic references were extracted using transformer-based NER models, cjber/reddit-ner-place_names (for place names) and Jean-Baptiste/roberta-large-ner-english (for POIs and landmarks). Gazetteer-based POI matching was also applied.

All extracted locations were geocoded using Nominatim (geopy) and clipped to the NSW boundary. A 5 km × 5 km grid was created, and both VIIRS hotspots and geocoded Twitter locations were overlaid to classify each grid cell into presence/absence categories. To test the association between hotspot locations and Twitter-derived locations, a Chi-square test was performed.

The study tested the below null-hypothesis and the alternative hypothesis.

Null hypothesis (H_0) - There was no association between the Twitter-derived locations and the hotspot locations, meaning both datasets were independent.
Alternative hypothesis (H_1) - There was an association between the two datasets.

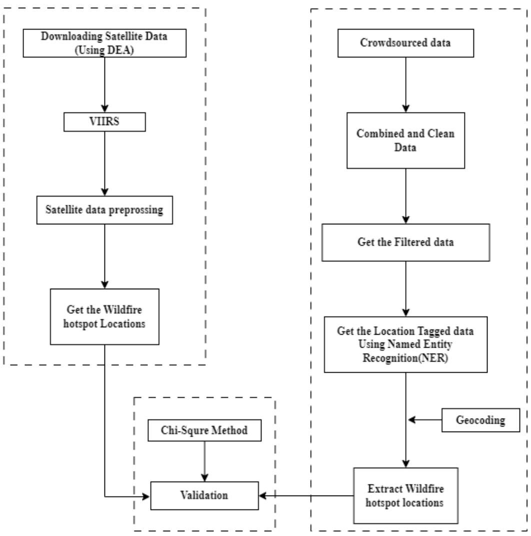

Fig. 1. Methodology overview

Expected frequencies were calculated using the Chi-square expected value formula:

$$E = \frac{(Row\ Total \times Column\ Total)}{Grand\ Total} \tag{1}$$

The Chi-square statistic was then computed using:

$$x^n = \sum_{i=1}^{n} \frac{(O_i - E_i)}{E_i} \tag{2}$$

where:
x^n = Chi-Square test statistic
O_i = Observed frequency in the i^{th} grid cell
E_i = Expected frequency in the i^{th} grid cell
n = Total number of grid cells

Degrees of freedom were calculated using the standard formula, and the chosen significance level (α) was used to obtain the critical Chi-square value. The computed Chi-square value was then compared with the critical value to decide whether to reject the null hypothesis.

Binomial tests checked if the association between Twitter locations and hotspots was positive or negative. Poisson and hypergeometric tests confirmed that observed overlaps (O_1) were higher than expected, validating the association's strength and direction.

III. Results and Discussion

During the study period, 12,178 wildfire hotspots were detected by the VIIRS sensor across NSW. Fig. 2 shows that hotspots were mainly concentrated in the Northeastern and Southeastern regions, while Central and Western NSW had fewer occurrences due to sparse vegetation.

The tweet dataset was cleaned for geocoding and spatial analysis by removing duplicates, missing values, URLs, and other unnecessary characters, and converting text to lowercase. Table 1 shows examples of raw versus cleaned tweets, ensuring a consistent and reliable dataset for further analysis.

After preprocessing, the dataset was filtered using wildfire-related keywords and hashtags linked to fire activity, emergency response, air quality, and geographic locations.

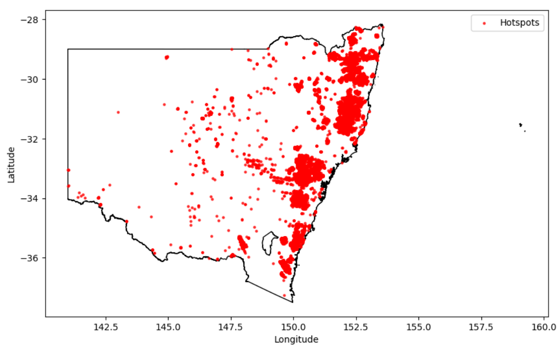

Fig. 2. Hotpots in NSW

Table 1. Examples of Raw and Cleaned Twitter Text After Preprocessing

Text	**Clean Text**
Long Gully Rd, Drake (Tenterfield LGA) Fire burning close to properties in Sandy Hill…	long gully rd drake tenterfield lga fire burning close to properties in sandy hill..
Much needed rainfall back home 🙏\n#MerryChrist...	much needed rainfall back home merrychristmas ...

The filtering process was highly effective, retaining most relevant tweets, and demonstrates the value of social media for disaster monitoring. However, the keyword-based approach may miss relevant tweets that use uncommon terms or slang, while some unrelated tweets may be incorrectly included due to ambiguous keywords.

Tweet activity rose from 1,648 in September to 39,112 in December (Fig. 3) reflecting increased public attention during severe wildfires.

Table 2 shows extracted locations and POIs from wildfire-related tweets using NER and gazetteer methods. Both broad areas (e.g., Clarence Valley) and specific sites (e.g., Wooli Street) were identified, demonstrating the method's ability to capture multi-level geographic information. This supports detailed spatial analysis for disaster response. Some noisy or duplicate entries remain, indicating the need for further refinement.

The cleaned dataset links each tweet ID to its final set of extracted locations. Table 3 shows that duplicates and generic terms were removed, leaving only meaningful place names and landmarks, making the dataset ready for spatial analysis.

The geocoding process converted extracted locations and POIs into latitude and longitude coordinates, linking each tweet reference to a spatial position (Table 4).

Nominatim provided reasonable accuracy for clear place names, but some ambiguous or landmark names resulted in errors.

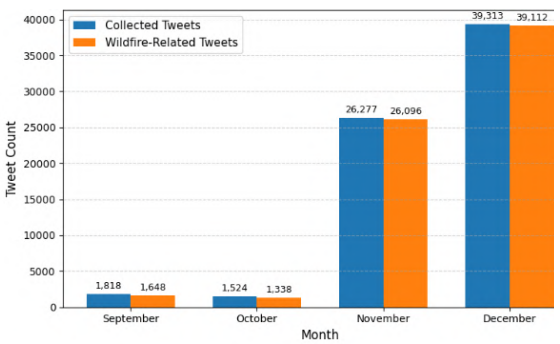

Fig. 3. Monthly Comparison of Collected vs Wildfire-Related Tweets

Table 2. Example of extracted locations and POIs from wildfire-related tweets

Clean Text	**Extracted Location**	**Extracted Landmarks and POIs**
emergency warning shark creek clarence valley lga crews are …	['clarence valley', 'shark creek', 'angourie',]	['yamba bowling club', 'wooli street']

Higher-precision services such as Google Maps, HERE, or Mapbox could improve accuracy, though they may involve costs and licensing restrictions.

Geocoded locations clipped to the NSW boundary, retaining only points within the study area (Fig. 4). Most points cluster in Eastern and Southeastern regions, particularly near coastal and urban zones, while inland and western areas show fewer points, reflecting lower population density and reporting frequency.

The NSW boundary was divided into a 5 × 5 km grid, with hotspots and geocoded Twitter locations. Both datasets cluster in eastern and southeastern regions, especially near coastal and urban areas, while inland and western areas have fewer points. This overlay demonstrates that Twitter data aligns with satellite-detected hotspots, capturing real fire activity at a local scale. Even a single tweet in a grid containing multiple hotspots indicates an accurate wildfire report within that 5 km cell.

The grid-based analysis of NSW (32,985 cells) showed that 3,150 cells contained data, with 396 cells having both Twitter locations and satellite hotspots, 951 cells with only Twitter data, and 1,803 cells with only satellite hotspots. This indicates that satellite data covered a larger area, while Twitter provided fewer but complementary points, and the overlap highlights where crowdsourced data supported satellite observations.

Table 3. Cleaned tweet IDs linked with extracted locations and POIs

Tweeter Id	Location or POI/Landmark
1.1713E+18	armidale
	tyringham
	dundarrah
1.17092E+18	clarence valley
	shark creek

Table 4. Results of Geocoded Location names

Location or POI/Landmark	Latitude	Longitude
armidale	-30.696454	152.211685
jeogla	-30.2152	152.1301374

The 396 overlapping cells are especially important because they show that some Twitter posts matched with satellite-detected hotspots. This highlights that the two data sources can complement each other and provide stronger evidence when combined. To check more formally whether the two datasets are related or independent, a Chi-square test of independence can be used.

Table 5 Contingency table created to summarize the relationship between satellite-detected hotspots and Twitter-derived locations.

The observed and expected frequencies for each category are shown in Table 6. These values were used to calculate the Chi-square statistic and test whether the two datasets are independent.

The Chi-square statistic (χ^2 = 1166.2851) was calculated using Eq (1).Based on the contingency table, the number of rows (r) was 2 and the number of columns (c) was also 2. Substituting these values into the formula gave:

$$df = (2 - 1) \times (2 - 1) = 1$$

Therefore, the degree of freedom was found to be 1.

And the significance level was set to $\alpha = 0.05$, which is the standard value used in chi-square hypothesis testing to ensure reliability of results.

From the chi-square table, the critical value for 1 degree of freedom at the 0.05 significance level was found to be 3.841. The calculated chi-square statistic was 1166.2851, which was much greater than the critical value of 3.841 at a 0.05 significance level.

$$({X^2}_{calculated}) >> ({X^2}_{critical})$$

Therefore, the null hypothesis was rejected, indicating that the Twitter-derived locations and satellite hotspots were not independent.

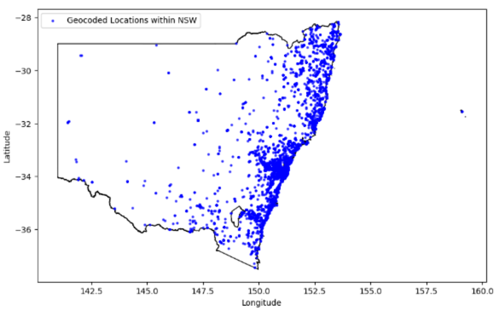

Fig. 4. Geocoded Locations within NSW boundary

Table 5 . Contingency table

	Tweeter Present	Twitter Absent	Total
Hotspot Present	396(O_1)	1,803(O_3)	2,199
Hotspot Absent	951(O_4)	29,835(O_2)	30,786
Total	1,347	31,638	32,985

The high Chi-square value comes mainly from the “both present” cell, where 396 grid cells had both Twitter locations and satellite hotspots, while the expected number under independence was only 89.80. The “both absent” cells also had higher observed values than expected. This shows that Twitter locations appear mostly in areas where fires actually happen and are missing in areas without fires.

Overall, the results of the Chi-square test provide strong evidence that the location accuracy of Twitter-derived data is high, making it a valuable resource for wildfire monitoring and disaster management applications.

However, to fully understand the nature of this association, whether it is positive or negative the study also performed seven additional probabilistic tests. These tests help to quantify the strength and direction of the relationship between Twitter-derived locations and satellite hotspots. Probability tests help us check if what we see is just by chance or if there is a real pattern. The observed number of hotspots and Twitter reports were compared with the expected facts if everything was random. The observed numbers are more different than the expected numbers, therefore the p-values are very small. This means the patterns are real and meaningful. The results of the seven probability tests show in Table 7. All the p-values are extremely small, which means the patterns are unlikely to happen by chance.

Overall, the results show a positive association between Twitter reports and satellite wildfire hotspots. This means that crowdsourced data from Twitter is generally accurate in terms of location and can be useful for monitoring wildfires in real time, supporting disaster management and emergency response.

Table 6. Observed (Oi) and expected (Ei) frequencies

	Tweeter Present		Twitter Absent	
	O_i	E_i	O_i	E_i
Hotspot Present	396	89.80	1,803	2109.20
Hotspot Absent	951	1257.20	29,835	29528.80

Table 7. Results of the seven probability tests

Test	p-value
P_1	3.85×10^{-125}
P_2	1.29×10^{-8}
P_3	9.49×10^{-13}
P_4	1.60×10^{-124}
P_5	1.97×10^{-4}
P_6	6.45×10^{-28}
P_7	5.45×10^{-153}

IV. CONCLUSION

This study reviewed the spatial accuracy of crowdsourced Twitter data to identify wildfire locations in New South Wales (NSW), Australia, during the 2019-2020 wildfire season by analyzing fire-related tweets with satellite detected hotspots of VIIRS. Here, data cleaning, NLP, NER, matching gazetteer and geocoding technique are the main components that used to develop the research and the integrated framework was created to combine social media data and satellite data in wildfire monitoring.

The results reveal that, after the proper processing, timely and detailed wildfire information are provided by Twitter data, which maintains strong spatial relationship with satellite identified hotspots. The significant positive correlation between the two datasets, illustrating that Twitter derived locations often reflect real wildfire activities, were confirmed by the statistical analyses including the chi-square test. While the satellite data offers expansive spatial coverage, social media contributes to improved local detail and faster updates in areas with active social media users.

The study also acknowledges several limitations. Geocoding accuracy was affected by ambiguous place names, informal language, and incomplete location references, which may have introduced positional uncertainty. In addition, social media reporting was higher in urban and coastal areas due to differences in population density, internet access, and user activity, potentially causing spatial reporting bias and underrepresentation in remote regions.

Another limitation is that the study mainly focused on spatial agreement between Twitter-derived locations and satellite hotspots, without detailed temporal analysis. Differences in tweet timing, satellite revisit intervals, the use of a single geocoding platform, and the fixed 5 km grid size may have influenced the observed spatial relationships.

As solutions, the framework by incorporating multiple geocoding services, higher resolution spatial validation methods, and temporal analysis should be improved for future research to evaluate the timing of wildfire reports better. The integration of additional crowdsourced platforms such as Facebook, Reddit, or emergency communication applications may also improve data coverage and reduce platform-specific bias. Furthermore, advanced transformer-based language models and multilingual NLP approaches enhances locational extraction accuracy from noisy social media text. In addition, machine learning approaches for credibility assessment and misinformation detection could strengthen the reliability of crowdsourced disaster information for operational use.

Overall, the study contributes to disaster management research by demonstrating the value of crowdsourced data, presenting methods for assessing spatial accuracy, and highlighting the role of NLP and geocoding in rapid disaster mapping. The findings support the combination of social media and satellite data for enhancing emergency response, situational awareness and real time decision making during a disaster event.

REFERENCES

[1] W. M. Jolly *et al.*, “Climate-induced variations in global wildfire danger from 1979 to 2013,” *Nat. Commun.*, vol. 6, Jul. 2015, doi: 10.1038/ncomms8537.

[2] T. Bell *et al.*, *FIRE IN THE LANDSCAPE FINAL PROJECT REPORT*. 2014.

[3] A. M. Graham *et al.*, “Impact of the 2019/2020 Australian Megafires on Air Quality and Health,” *Geohealth*, vol. 5, no. 10, Oct. 2021, doi: 10.1029/2021GH000454.

[4] Catherine. Stephenson, *A literature review on the economic, social and environmental impacts of severe bushfires in south-eastern Australia*. Victorian Government Department of Sustainability and Environment, 2010.

[5] D. M. Nelson, Y. He, and G. W. K. Moore, “Trends and applications in wildfire burned area mapping: Remote sensing data, cloud geoprocessing platforms, and emerging algorithms,” Jul. 01, 2024, *Elsevier B.V.* doi: 10.1016/j.geomat.2024.100008.

[6] Y. Zhao and Y. Ban, “Near real-time wildfire progression mapping with VIIRS time-series and autoregressive SwinUNETR,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 136, Feb. 2025, doi: 10.1016/j.jag.2025.104358.

[7] E. S. Hope *et al.*, “Exploring the use of satellite Earth observation active wildland fire hotspot data via open access web platforms,” *Int. J. Digit. Earth*, vol. 17, no. 1, 2024, doi: 10.1080/17538947.2024.2420821.

[8] Y. Chen, D. C. Morton, and J. T. Randerson, “Remote sensing for wildfire monitoring: Insights into burned area, emissions, and fire dynamics,” Jun. 21, 2024, *Cell Press*. doi: 10.1016/j.oneear.2024.05.014.

[9] A. Sathiyamoorthy, W. Zhou, X. Zhou, X. Li, and I. Gondal, “MFiSP: A Multimodal Fire Spread Prediction Framework,” Oct. 2025, [Online]. Available: http://arxiv.org/abs/2510.23934

[10] C. Songchon, G. Wright, and L. Beevers, “Quality assessment of crowdsourced social media data for urban flood management,” *Comput. Environ. Urban Syst.*, vol. 90, Nov. 2021, doi: 10.1016/j.compenvurbsys.2021.101690.

[11] X. Zhong, M. Duckham, D. Chong, and K. Tolhurst, “Real-time estimation of wildfire perimeters from curated crowdsourcing,” *Sci. Rep.*, vol. 6, Apr. 2016, doi: 10.1038/srep24206.

[12] S. Sachdeva and S. McCaffrey, “Using social media to predict air pollution during California wildfires,” in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Jul. 2018, pp. 365–369. doi: 10.1145/3217804.3217946.

[13] A. Mehmood, M. T. Zamir, M. A. Ayub, N. Ahmad, and K. Ahmad, “A Named Entity Recognition and Topic Modeling-based Solution for Locating and Better Assessment of Natural Disasters in Social Media,” May 2024, [Online]. Available: http://arxiv.org/abs/2405.00903

[14] Y. M. Chang, S. Rakshit, C. H. Huang, and W. H. Wu, “Probabilistic approaches for investigating species co-occurrence from presence-absence maps,” *PeerJ*, vol. 11, 2023, doi: 10.7717/peerj.15907.