Social media sentiment analysis and opinion mining in public security: Taxonomy, trend analysis, issues and future directions

1. Introduction

- Ensuring public security is crucial for maintaining stability in a country, expanding beyond traditional notions to include environmental, societal, economic, and political dimensions. The concept of security now extends to individual safety and well-being, emphasizing protection against both criminal and non-criminal threats. Public security involves maintaining social privacy, eliminating risks, and optimizing opportunities for sustainable development and well-being.

1.1 Public Security:

- Broadened Scope: Security encompasses protection against attacks, danger, and the state of feeling happy and safe.
- Components of Public Security: Involves maintaining social privacy, eliminating risks, and optimizing opportunities for sustainable development and well-being.

1.2 Role of Opinion Mining (Sentiment Analysis) in Public Security:

- Evolution of Security Definition: Reflects the changing landscape of security, including individual well-being.
- Threats to Public Security: Criminal and non-criminal threats, such as terrorism, riots, protests, crises, accidents, crime, natural disasters, disease outbreaks, and pandemics.
- Impact of Events: Significantly threatens human life, safety, and causes economic and cultural losses.

1.3 Opinion Mining and Sentiment Analysis:

- Definition: Field analyzing opinions, sentiments, attitudes, and emotions in written text.
- Abstraction Levels: Opinion mining deals with concrete views, while sentiment analysis focuses on attitudes prompted by feelings.

- Applications: Originally used for product reviews, expanded to stock markets, elections, disasters, healthcare, and software engineering.

1.4 Sentiment Analysis in Public Security:

- Shift in Applications: Originally for product reviews, now applied to analyze sentiments and public opinions during disasters, emergencies, and events.
- Data Source: Utilizes social media data for sentiment analysis due to its vastness and real-time nature.

1.5 Purpose of the Survey:

- Overview: Provides an overview of sentiment analysis and opinion mining in public security.
- Taxonomy Development: Aims to develop a descriptive taxonomy based on recent research.
- Analysis of Trends: Explores current trends in sentiment analysis and opinion mining.
- Identification of Issues: Identifies current issues and suggests potential future research directions.

2. Methodology

2.1 Database Selection:

- Criteria: Databases selected based on wide coverage of scientific peer-reviewed articles and strict journal evaluations.
- Selected Databases:
    - Scopus
    - IEEE Xplore
    - Science Direct

2.2 Keyword Search:

- Search Terms: Various combinations of terms and operators used, including “sentiment AND analysis AND public AND security,” “opinion AND mining AND public AND security,” and other relevant variations.
- Search Results:
    - Scopus: 2097 articles

- IEEE Xplore: 669 articles
- Science Direct: 239 articles

2.3 Screening Process:

- Removal of Non-Academic Articles: Non-academic articles were excluded from consideration.
- Article Types Considered:
  - Journals
  - Conference Proceedings
  - Serials
- Duplication Removal: Ensured that duplicate articles across databases were removed.
- Publication Year Criteria: Included only articles published in recent years (2016–2023).
- Remaining Articles After Screening:
  - Scopus: 1903 articles
  - IEEE Xplore: 663 articles
  - Science Direct: 166 articles

2.4 Eligibility Filtering Process:

- Sub-Processes:
  - (i) Combination of databases and removal of duplicates
  - (ii) Brief review of article titles, abstracts, and keywords
- Number of Articles After Filtering:
  - Reduced to 1485 articles after removing duplicates
- Manual Review Criteria:
  - Identification of articles employing sentiment analysis/opinion mining in the public security domain
  - Articles written in English
- Final Selection:
  - 280 articles were selected for a thorough review

2.5 Thorough Review and Eligibility Confirmation:

- Content Review: Thorough examination of the selected 280 articles.
- Eligibility Confirmation:
  - Based on criteria, 200 articles were deemed eligible for inclusion in this paper.

3. An Overview of Sentiment Analysis and Opinion Mining for Public Security

3.1 Data Acquisition:

- Datasets are acquired from social media platforms or other relevant sources using keywords, geographic location information, or specific timeframes.
- Publicly shared datasets may be used, eliminating the need for data acquisition.

3.2 Pre-processing:

- Raw data undergoes pre-processing using text-processing techniques.
- Techniques include text cleaning, normalization, replacement, and stopword removal.
- Pre-processing aims to remove noise and irrelevant data, preparing for the feature engineering stage.

3.3 Feature Engineering:

- Involves feature extraction, selection, and representation.
- Features are extracted from pre-processed data, representing the original text in a numerical form compatible with algorithms.
- Techniques include statistical, NLP, rule-based methods, and deep learning for learning multiple levels of representation.

3.4 Sentiment or Opinion Classification:

- Classification algorithms, often founded on a lexicon, are used.
- Lexicon-based approach uses sentiment resources like lexicons or corpus databases.
- Sentiment scores are calculated and evaluated based on sentiment orientation and strength.
- Topic modeling is used in some approaches to categorize topics within the dataset.

3.5 System Performance Evaluation:

- Evaluation metrics depend on the approach adopted for analysis.

- Machine learning-based approaches use metrics like accuracy, precision, recall, and F1-measures.
- Lexicon-based approaches often use accuracy based on sentiment score calculation.

3.6 Result Analysis:

- Sentiment analysis results provide insight into the event of interest.
- Analysis may predict future occurrences or offer a retrospective analysis of the event.

This comprehensive framework serves as a foundation for understanding sentiment analysis and opinion mining in public security. The subsequent sections will delve into specific sub-branches and trends within this domain, providing a detailed taxonomy and analysis.

4.1. Objective of Sentiment Analysis and Opinion Mining:

- **4.1.1. Analysis of Events:**
    - Focus on disease outbreaks and pandemics (e.g., COVID-19, MERS, Zika virus).
    - Objectives include sentiment analysis and topic modeling.
    - Output: Analytical studies on outbreak identification, pandemic tracking, and statistical analysis.
- **4.1.1. Disaster or Emergency Management:**
    - Aimed at improving disaster management and exploring emergency events.
    - Output includes sentiment and disaster concern index, event urgency level, and quantitative analysis of public resilience.
    - Decision support systems for disaster management are explored.
- **4.1.2. Improvement of Techniques:**
    - Feature Engineering:
        - Focus on feature enhancement and representation techniques.
        - Information enrichment involves geographical and location data.
        - Lexicon expansion aims to improve vocabulary and dictionary for evolving language.
        - Output: Methodology for sentiment analysis from a geographic perspective and spatio-temporal approaches.
    - Classification Techniques:
        - Modification of classifiers using machine learning, deep learning, and hybrid ensembles.

        - Techniques for annotation, pipeline, parsing, and lexicon utilization explored.
        - Output: Techniques, models, frameworks, and algorithms for sentiment analysis in public security.
    - Event Prediction Technique:
        - Focus on improving techniques for detecting and predicting events.
        - Objectives include early event detection and real-time monitoring.
        - Output: Improved approaches for early detection and prediction of emergency/disaster events.
- **4.1.3. Corpus Generation:**
    - Initiation of Corpus:
        - Creation of new corpus or addition to existing corpus.
        - Compiled corpora concerning specific events related to public security.
    - Technique Enhancement:
        - Refinement and enhancement of existing techniques for better data curation.
        - Output includes annotated corpora, data collection methodologies, and dataset manipulation.

4.2. Domain of Interest in Public Security:

- **4.2.1. Natural:**
    - Events triggered by natural causes (e.g., earthquakes, floods, pandemics like COVID-19).
    - Studies on single and combined events, exploring multiple categories.
- **4.2.2. Non-natural (Human-made):**
    - Events caused by humans, including emergencies, chaos, crises, and cyber/technology threats.
    - Encompasses terrorism, riot, protest, crime, conflicts, immigrant issues, economic crises, and cyber-attacks.

4.3. Public Security Event Timeframe:

- **4.3.1. Pre-event:**
    - Least explored timeframe involving sentiment analysis before event occurrence.
    - Potential for predicting threats and implementing preventive measures.
- **4.3.2. During-event:**
    - Real-time sentiment analysis during ongoing events.
    - Valuable for time-sensitive issues with a large social media following.

- **4.3.3. Post-event:**
    - Most prevalent type of study involving sentiment analysis after an event.
    - Aids in re-examining issues, supporting policy-making, and providing insights for future preparations.

4.3. Public Security Event Timeframe

- **4.3.1. Pre-event**
    - Least explored timeframe in literature.
    - Potential for predicting and preventing events.
    - Abid et al. (2017): Proposed methodology to predict potential threats based on social media content related to terrorism.
    - Almehmadi et al. (2017): Used Twitter data to predict crime rates in Houston and New York City.
- **4.3.2. During-event**
    - Second most common area of study in social media sentiment analysis.
    - Provides real-time insights into public sentiment during ongoing events.
    - Useful for time-sensitive issues with large social media followings.
    - Studies, like those on disease outbreaks and the coronavirus pandemic, analyze data collected within specific timeframes.
    - Some studies include data collected prior to the known occurrence of an event for comprehensive analysis.
- **4.3.3. Post-event:**
    - Post-event sentiment analysis and opinion mining are prevalent in public security research.
    - Analysis involves examining data after a specific event to re-examine issues and provide insights.
    - Aims to support policy making, management, and preparation for future endeavors.
    - Examples include immigration breaches, shooting incidents, collision accidents, nuclear accidents, and natural disasters.

4.4. Social Media Platform:

- Recent work primarily uses either a single social media platform or a mix of multiple platforms.
- Twitter and Sina Weibo are commonly used, followed by YouTube, Facebook, and other microblogs.

- Some studies combine platforms like Twitter and Facebook, Twitter and YouTube, Twitter and Flickr, etc.
- Twitter is favored for its real-time information dissemination, while platforms with data acquisition restrictions, like Facebook, are less popular.

4.5. Dataset:

- The type of dataset used falls into two categories: public and private.
- Public datasets, available in the public domain, allow researchers to compare results and aid in transfer learning.
- Examples of public datasets include geotagged tweets related to disasters, COVID-19-related datasets, and datasets on specific events like the refugee crisis, Palestinian-Israeli conflict, Sewol Ferry Disaster, terrorism, and cyber security.

4.6. Language of Dataset:

- The language attribute categorizes datasets into monolingual, bilingual, and multilingual.
- Monolingual datasets use a single language, with English being the most common, followed by Chinese and other languages.
- Bilingual datasets combine two languages, often involving translation.
- Multilingual datasets encompass three or more languages, with studies focusing on code-switched or code-mixed datasets, such as Malay-English, Filipino-English, Indonesian-English, and Nigerian-English datasets.

4.7. Sentiment Analysis and Opinion Mining Approach:

- **4.7.1. Machine Learning-Based Approach:**
    - Supervised Learning:
        - Utilizes labeled data for sentiment classification.
        - Algorithms: Random Forest, Support Vector Machine, Naïve Bayes, K-Nearest Neighbor, Logistic Regression, Maximum Entropy.
        - Challenges: Domain-dependency, effort in creating training sets.
    - Unsupervised Learning:
        - Doesn't require labeled data; uses hidden structures.
        - Approaches: Clustering (K-Means, K-Medoids, DBSCAN), Topic Modeling (Latent Dirichlet Allocation).
        - Challenges: Unreliable for non-dependent features, out-of-domain keywords.

    - Semi-Supervised Learning:
        - Uses small labeled data and a larger unlabeled set.
        - Challenges: Lower classification performance, not suitable for rule-based feature extraction.
    - Ensemble Learning:
        - Combines outputs for improved classification (bagging, boosting, stacking, RF).
        - Boosting commonly outperforms bagging.
        - Requires a deeper understanding of the dataset.
    - Deep Learning:
        - Subset of machine learning using deep neural networks.
        - Examples: Bi-Directional LSTM, CNN, LSTM, BERT, ERNIE.
        - Advantages: Better classification with deep features.
        - Disadvantages: Large data requirements, high training costs.
- **4.7.2. Lexicon Approaches:**
    - Dictionary-Based:
        - Uses a dictionary with word polarities (VADER, WordNet, SentiWordNet).
        - Simple implementation, doesn't need labeled data.
        - Challenges: Limited by dictionary quality and coverage.
    - Corpus-Based:
        - Relies on co-occurrence patterns and seed word lists.
        - Requires a large training corpus.
        - Provides better performance in specific domains.
- **4.7.3. Hybrid Approach:**
    - Combines machine learning and lexicon-based approaches to compensate for individual shortcomings.
    - Uses lexicons for feature extraction and machine learning for sentiment classification.
    - Improved performance compared to individual approaches.
- **4.7.4. Manual Coding:**
    - Relies on human coders to read and code datasets using pre-defined rules.
    - Specific rules for positive and negative statements.
    - Advantages: Produces better performance, provides deeper insights.
    - Disadvantages: Time-consuming, not easily transferable to other domains.

5. Analysis of trend, issues and future directions of sentiment analysis and opinion mining for public security

5.1. Trend Analysis of Sentiment Analysis Objectives:

- Objective Trends:
  - Analysis of events and technique improvement are primary research objectives, showing an upward trend since 2019.
  - Multipurpose objectives slightly increased since 2020, while corpus generation remains limited.
- Distribution of Objectives (2016-2023):
  - Analysis of events and technique improvement account for 41% and 39% of the work.
  - Multipurpose objectives constitute 16%, while corpus generation is the least common objective.

5.2. Trend Analysis of Public Security Domains:

- Domain Trends:
  - Natural disaster domain increased until 2020, then decreased, while the public health domain surged from 2020.
  - Non-natural domain of emergency increased from 2019 to 2020, decreased in 2021 and 2022, with chaos and multi-domain interests increasing.
- Distribution of Recent Work:
  - Public health and natural disasters dominate recent work at 32% and 25%.
  - Emergency domain follows at 16%, chaos at 13%, and crisis at 5%.
  - Cyber and technology domain has the second-lowest percentage at 3%.
  - Integration of natural disaster and public health domains is limited (2%).

5.3. Trend Analysis of Event Timeframe:

- Event Timeframe Trends:
  - Post-event timeframe attracted most interest until 2020, when during-event timeframe saw a significant increase due to the COVID-19 pandemic.
  - Pre-event timeframe interest remains consistently low.
- Distribution of Published Work Across Timeframes:
  - During-event and post-event timeframes have high publication rates at 50% and 36%.

    - Multiple and pre-event timeframes have lower rates at 10% and 4%, indicating a focus on addressing post-event needs rather than prediction or early detection.

5.4. Trend Analysis of Social Media Platforms:

- Twitter and Sina Weibo are popular for public security research due to easy data acquisition.
- Mixed platforms show a slight decrease in popularity, possibly due to user preferences.
- Twitter dominates sentiment analysis, followed by Sina Weibo; other platforms are underutilized.

5.5. Trend Analysis of Language of Dataset:

- Monolingual datasets are increasingly used, with a growing interest in bilingual and multilingual data.
- Bilingual studies address unique language challenges in sentiment analysis.

5.6. Trend Analysis of Dataset Type:

- Private datasets are more prevalent (77%), with a steady increase.
- Public datasets see a slight increase, primarily driven by the availability of COVID-19 datasets.
- Public health and natural disaster datasets are most utilized.

5.7. Trend Analysis of Approaches for Sentiment Analysis:

- Dictionary-based approach is most preferred, with a steady increase.
- Hybrid and deep learning approaches gain popularity since 2020.
- Twitter remains the dominant platform, with 59% of published work.

5.8. Issues and Future Directions:

- **5.8.1. Issues:**
    - Shortage of multi-class and different-level analysis approaches.
    - Insufficient availability of public security domain-independent datasets.
    - Inadequate prediction based on timeframe coverage.
    - Lack of supporting approaches for variations across different languages.
    - Limited information availability.
- **5.8.2. Future Works:**
    - More extensive multi-class and multi-level analysis.

- Production of multi-domain public security datasets.
- Utilization of a greater variety of features and techniques for automatic detection of outbreaks.
- Establishment of cross-language corpora and supporting approaches.
- Expansion of data acquisition and geographical coverage.