IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 6, 2021, accepted February 28, 2021, date of publication March 4, 2021, date of current version March 16, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3063921

A Survey on Opinion Reason Mining and Interpreting Sentiment Variations

FUAD ALATTAR AND KHALED SHAALAN
Faculty of Engineering and IT, The British University in Dubai, Dubai, United Arab Emirates
Corresponding author: Fuad Alattar (fuad.alattar@hotmail.com)

ABSTRACT Tracking social media sentiment on a desired target is certainly an important query for many decision-makers in fields like services, politics, entertainment, manufacturing, etc. As a result, there has been a lot of focus on Sentiment Analysis. Moreover, some studies took one step ahead by analyzing subjective texts further to understand possible motives behind extracted sentiments. Few other studies took several steps ahead by attempting to automatically interpret sentiment variations. Learning reasons from sentiment variations is indeed valuable, to either take necessary actions in a timely manner or learn lessons from archived data. However, machines are still immature to carry out the full Sentiment Variations' Reasoning task perfectly due to various technical hurdles. This paper attempts to explore main approaches to Opinion Reason Mining, with focus on Interpreting Sentiment Variations. Our objectives are investigating various methods for solving the Sentiment Variations' Reasoning problem and identifying some empirical research gaps. To identify these gaps, a real-life Twitter dataset is analyzed, and key hypothesis for interpreting public sentiment variations are examined.

INDEX TERMS Emerging topic, event detection, interpreting sentiment variations, opinion reason mining, sentiment analysis, sentiment reasoning, sentiment spikes, topic modeling.

I. INTRODUCTION

Sentiment Analysis – also known as Opinion Mining – is a Natural Language Processing (NLP) task that received a lot of attention during the last two decades due to the necessity of automatic review of social media, news websites, blogs, etc. This analysis became crucial for those who are interested in monitoring users' feedback about specific targets like products, events, public entities, etc. Such feedback is essential for decision-makers who need to take necessary actions based on public reactions.

E-Marketing is a good example of applications that employ Sentiment Analysis techniques. Automatic generation of marketing material may target social media users based on tracking their online feedbacks. Positive feedback about products or product-features can be used as triggers for online advertising, whereas negative feedback may help manufacturers in taking necessary corrective actions for the production process. Another application of Sentiment Analysis is polls on politicians. Social media can be considered as a dashboard that reflects public acceptance/rejection of certain policies or politicians. Therefore, early sensing of these sentiments through online sources may help these public entities to adapt their policies and actions accordingly.

Many other applications have utilized Sentiment Analysis techniques. You may refer to article [1] for more examples. Most of Sentiment Analysis studies - like [2]–[5] - had focused on identifying subjectivity and polarity of texts, either on document-level or sentence-level. However, these techniques do not indicate possible motives behind consumers' opinions. As a result, some researchers tackled the Reason Mining task to get the best out of the Sentiment Analysis exercise as presented in Sections II & III of this survey.

In this survey, the term "Reason Mining" is chosen for these tasks in which feedback texts are monitored to conclude possible reasons that caused either extracted sentiment itself or major changes in sentiment levels. These reasons could be some product features, political decisions, rumors, national disasters, etc.

In their study of current challenges and new directions of Sentiment Analysis research, Poria *et al.* [6] acknowledged the high importance of the Sentiment Reasoning task. Moreover, they prophesized that this task will be one of the main future directions of Sentiment Analysis field. Refer to Fig. 1 for the expected directions of Sentiment Analysis research.

The associate editor coordinating the review of this manuscript and approving it for publication was Xinyu Du.

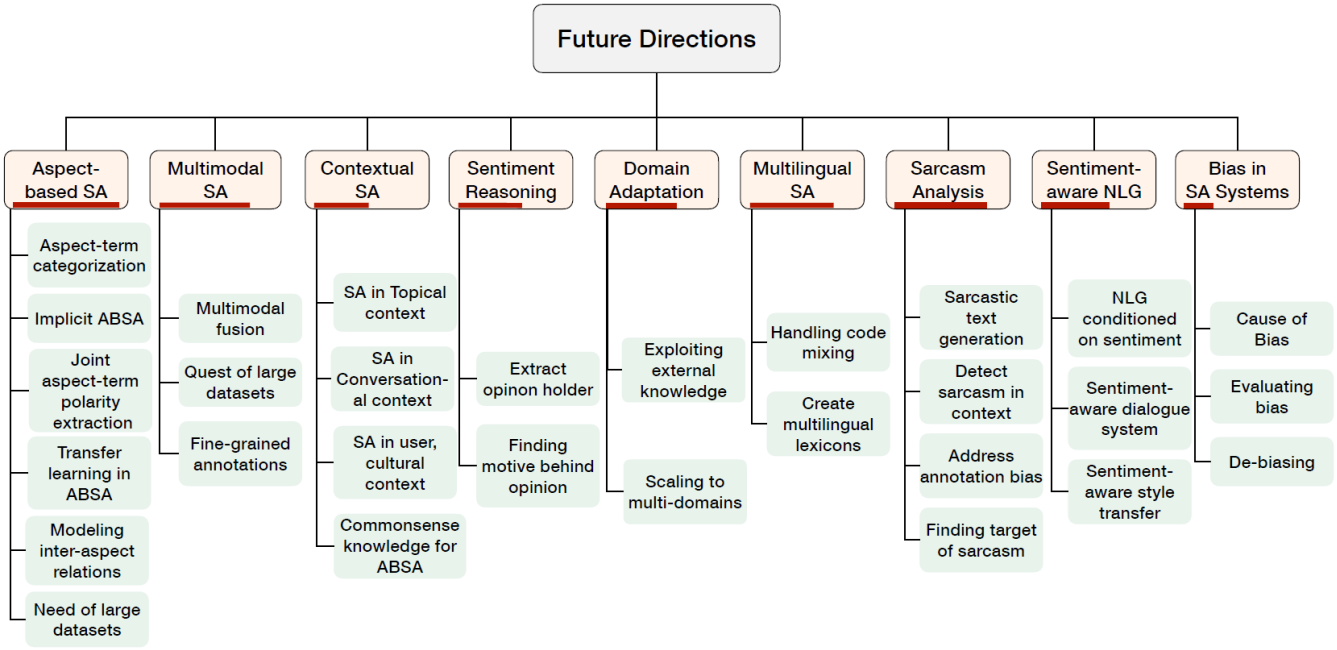

FIGURE 1. Future directions of sentiment analysis research [6].

A. RESEARCH QUESTIONS

The focus of this survey is the Sentiment Variations' Reasoning problem, and its application on Twitter. Our objectives are to qualitatively examine various methods for solving this problem and discover some research 'Empirical Gaps' [7] and ''Evaluation Voids'' [8]. After exploring these methods, main hypothesis used in relevant studies are discussed. These main research questions (RQ) shall be addressed:

RQ (1) Explicit Reasons: Do most of subjective tweets explicitly indicate reasons of sentiment?

RQ (2) Aspects: Can Aspect-Based method always capture reasons of sentiment in products and services domains?

RQ (3) Topic Frequency: In a sentiment variation spike, does the main reason of the spike always have the highest topic frequency?

RQ (4) Events: Can the Event Detection method always discover reasons of public sentiment variations?

RQ (5) Emerging Topic: Is the Emerging Topic detection efficient for capturing sentiment variations?

RQ (6) Topic Visualization: Can Topic Visualization enhance our understanding of topic evolution inside a document set?

RQ (7) Sentiment Spike: Does a sentiment variation reason always cause a spike in the overall sentiment level?

RQ (8) Topic Modeling: Can practical Topic Modeling methods help us understand reasons of sentiment variations?

RQ (9) Foreground-Background Topics: Can the FB-LDA Model discover Emerging Topics within a sentiment variation period?

B. SURVEY OUTLINE

Sections II and III investigate main Reason Mining approaches in the field of Sentiment Analysis by selecting representative articles for each approach. Fig. 2 summarizes these approaches.

Section IV. aims to discuss the Sentiment Variations' Reasoning problem and attempt to answer main research questions through practical experiments that use two annotated datasets. Finally, conclusions are presented in Section V.

II. SENTIMENT REASONING

This section is a selective literature review where the first branch of the Sentiment Reason Mining problem is explored, and the methods for discovering reasons behind expressed sentiment are reviewed. Though our survey focuses on the second branch of the Sentiment Reasoning problem – i.e., Sentiment Variations Reasoning - we find it helpful to explore the methods of extracting reasons behind sentiment itself because these reasons may contribute to the task of interpreting sentiment variations.

Representative articles are selected in this section to help us understand Sentiment Reasoning approaches; however, our survey did not attempt to address all written articles that contributed to Sentiment Reasoning studies. Furthermore, given that some of the addressed approaches are used for many other NLP and Machine Learning tasks, we did not make a deep dive into sub-categories of each approach as these are not explicitly related to Reason Mining studies.

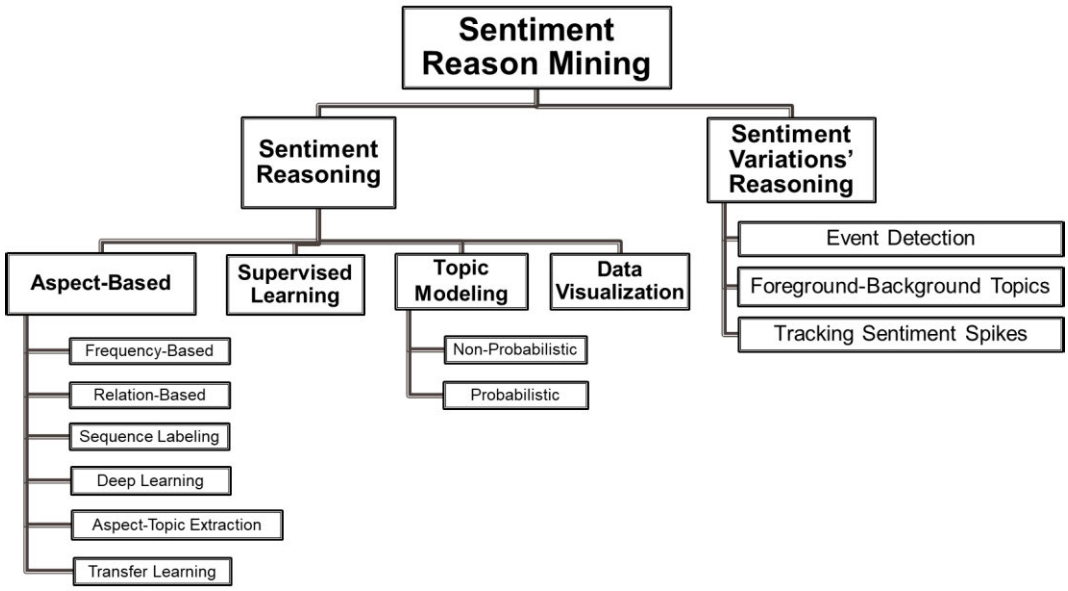

FIGURE 2. Approaches of reason mining for sentiment analysis.

To illustrate, Topic Modeling methods are categorized here into Non-Probabilistic and Probabilistic, however the Supervised, Unsupervised and Semi-supervised techniques of both Topic Modeling methods are not addressed in this survey as these are also used in many other applications where Topic Models are employed.

A. ASPECT-BASED SENTIMENT ANALYSIS

Aspect-Based Sentiment Analysis (ABSA) is a sub-field of Sentiment Analysis that aims to extract opinions on specific features or aspects of the desired target [9]. It is also known as Aspect-Oriented, Aspect-Level, Feature-Based and Feature-Oriented Sentiment Analysis. For instance, if our target is a printer, the ABSA extracts users' sentiment about features like speed, ink type, noise levels, power consumption, cost, etc.

By tracking users' sentiment about various aspects/features of our target, the reasons of public sentiment about the target itself could be understood when these sentiments are linked to one or more aspects of that target [10]. If our target is a printer's brand name, then customers' dissatisfaction about a certain feature like ''noise level'' could be the main reason for a highly negative sentiment level on printer's brand name itself.

Aspect of the target can be either Explicit or Implicit, where extraction of the later is more challenging [11]. To illustrate, the sentence ''This mobile is too bulky to put in my pocket'' contains an Implicit aspect because it addresses the ''Size'' aspect without explicitly mention the word ''Size''. In addition to this challenge, Aspect-Based method is sensitive to domain changes. For example, the word ''Hot'' could be a good feature for some products, whereas it should be considered negative when it describes products like batteries.

Given that the reasons behind sentiment can be something else other than the features of the target, the Aspect-Based method cannot capture sentiment reasons that are not categorized as aspects. To show you what we mean, a corruption scandal related to a manufacturing company may impact the sentiments towards its products regardless how good their features are. Moreover, it is difficult to apply Aspect-Based methods on targets outside the products and services domains. If the target is an event, it would be almost impossible to extract related features/aspects using standard ABSA methods [12].

1) FREQUENCY-BASED

Some studies adopted the Frequency-Based approach to extract aspects from text. This technique extracts ''explicit aspect expressions'' which are noun-phrases and nouns from a large-scale review data. Hu and Liu [13] worked on opinion summarization for customer reviews, which gives the user an overview about main reasons behind certain sentiments towards the desired product. Their work mines the product's aspects that received positive/negative feedbacks from reviewers. The following steps are followed to achieve the opinion summarization task: Identify the aspects of the target, detect sentiments inside each customer-feedback, and summarize all results.

Targeting customer reviews on five different products, Hu and Liu [13] could achieve an average accuracy of 84% for identifying the polarity of sentiment. However, their work faced some limitations as they could not handle opinions that require pronoun-resolution. Their technique could not recognize what a pronoun in a sentence represents or refers to. For instance, the pronoun ''it'' in the sentence ''It is efficient and fast'' could not be analyzed although it refers to the target product/feature. Furthermore, their technique

focused on adverbs only and ignored opinions which are expressed through verbs and nouns. For instance, the polarity of the sentence ''I love this car'' cannot be detected by their developed model.

Popescu and Oren [14] enhanced the approach of [13] by using OPINE method which analyzes product reviews to construct a model of product features and their assessment by reviewers. OPINE tries to take out all nouns that are not aspects or entities, and it employs an unsupervised learning technique. This method could obtain 22% higher accuracy when compared to the results of [13].

To enhance the Frequency-Based method further, Scaffidi *et al.* [15] compared the frequency of detected nouns and noun-phrases with their frequency in English-Corpus. Blair-Goldensohn, *et al.* [16] also aimed to enhance the method by considering the nouns which are included in subjective sentences.

O'Connor *et al.* [17] analyzed multiple surveys/polls on consumer and political related issues for both years 2008 and 2009. They observed that 80% of these surveys correlate to the frequency of used sentiment words inside tweets. However, they used a simple detector to extract sentiments which could not handle noisy Twitter data. The Frequency-Based technique was refined further in [18] by employing a syntactic pattern-based filter to remove terms which are not aspects/features.

In general, the Frequency-Based method suffers the limitation of missing out low-frequency aspects, in addition to its need for manual configuration and tuning of parameters to suite the selected dataset [19].

2) RELATION-BASED

The Relation-Based approach, also known as Rule-Based and Syntax-Based [20], tries to find the relation between target's features and sentiment words to extract the aspects. To illustrate, the expression ''Awesome brightness'' represents an adjectival modifier relation between the adjective ''Awesome'' and the aspect ''Brightness''. Zhuang *et al.* [21] used the Relation-Based method to extract features from customer reviews. They employed a dependency-parser to detect the relation between features and sentiment words.

Wu *et al*. [22] enhanced the Relation-Based method by using a phrase-dependency-parser to extract aspects which are noun-phrases or verb-phrases. Qui *et al.* [23] employed a double-propagation technique to identify relationships between opinion words and aspects.

The authors of [24] focused on classifying Twitter ''target-dependent'' sentiments. They noticed that previous approaches employ ''target-independent'' algorithms and processes, which may cause the system to identify sentiments that are not relevant to the target. They also noticed that these approaches address each tweet separately and do not consider other related tweets. As a result, they developed a process to resolve this limitation. They first categorized the tweets into positive, negative or neutral/objective based on the detected opinions of the tweets. In their research, they considered the input ''query'' to be the target of the sentiment. They also considered related tweets when they classified sentiment using ''graph-based'' optimization. According to the published results, their method showed higher performance when compared to target-independent approaches. However, their study suffers a genuine limitation because it considers all noun phrases related to the target as ''extended-targets''. This may mistakenly link irrelevant sentiment to the target. Assume that the engine of a car is identified as an ''extended target'', then a negative sentiment about the engine can be extended to the car itself, which is acceptable; however, if the same is applied to the engine's oil of the car, a false sentiment about the car could be concluded by applying a sentiment that is related to its engine's oil only.

Unlike Frequency-Based approach, Relation-Based approach can detect low-frequency aspects, however More and Ghotkar [25] argued that this approach may produce many terms which are not real aspects. It is also confirmed in [26] that Relation-Based approach may extract irrelevant features.

3) SEQUENCE LABELING

Sequential Labeling methods like Hidden Markov Model (HMM) and Conditional Random Field (CRF) were used in a supervised-learning mode to extract aspects of the target. In [27], HMM is used in a framework that integrates linguistic elements like Part-Of-Speech (POS) into a learning process to predict patterns and correlations between tags. In [28], a CRF model is used to detect boundaries of sentiment phrases and identify both polarity and intensity of these phrases.

A Hierarchical Sequence Labeling Model (HSLM) is developed in [29]. It consists of three elements: aspect-level, opinion-level, and sentiment-level. The model learns the interactions between these three elements through a special information fusion technique.

It is explained in [26] that - generally - the sequence labeling techniques suffer the limitation in handling dependencies between multiple labels, hence they are unable to capture the complete meaning of the sentence. In [30], Sequence-to-Sequence (Seq2Seq) learning is used to reduce the impact of this limitation. Seq2Seq considers relations between the opinion polarity of features in feature-level opinion classification method.

4) DEEP LEARNING

Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and other Neural Networks are widely used for various Sentiment Analysis tasks, including extraction of aspects. In [31], RNN is used with Word Embedding to develop a discriminative model for Aspect-Extraction. Similarly, RNN is used in [32] for the Two-Step Aspect-Extraction method. RNN is also applied in [33] for the Financial domain on tweets and news headlines. CNN is used in both [34] and [35] to carry out aspect-extraction as a multi-label classification problem,

whereas it is combined with RNN in [36] to handle Aspect-Extraction and Sentiment Analysis tasks.

The mechanism of Attention Models was introduced to enhance the performance of Deep Learning methods. Zhang and Lu [37] developed a Multi-Attention Network to handle the Aspect-Extraction task.

It is indicated in [38] that model over-fitting is a main challenge for Deep Learning techniques. This becomes more visible when training dataset has limited number of domains. Although it is acknowledged in [39] that Deep Learning techniques have achieved good accuracy levels, it is also mentioned that these algorithms need high computation power because of their complexity. These techniques also suffer the ''opacity problem'' because it is not always clear how Deep Learning models make their decisions after being trained.

5) ASPECT-TOPIC EXTRACTION

Probability distribution over words is known as ''Topic''. To simplify the concept of topics, assume that you got a bag of papers where each paper contains a word. Probability of pulling each word from the bag is greater than zero. If the bag contains two papers with the word ''Dubai'' and one paper with the word ''Paris'', then the chance of pulling the word ''Dubai'' is 0.666, whereas 0.333 is the chance of pulling the word ''Paris''. This example illustrates the concept of a topic. It gives us the probabilities/chances of a set of words for the assigned topic, and it is the basis of Topic Modeling methods [40].

To illustrate, the Topic Model of Latent Dirichlet Allocation (LDA) [41] considers each document as a mix of topics which exist in the corpus. The model suggests that each word in the document is linked to one of the topics inside the document. For instance, if LDA output gives word probabilities of 45% for the word ''Restaurants'', 10% for the word ''Transport'', and 45% for the word ''Hotels'', this output indicates that the discussed topic in the selected document is ''Facilities'' [40]. The Topic Model deals with a document set as a Bag of Words, therefore it neither considers the order of the words nor the grammar of the sentences.

The Aspect-Topic Extraction method utilizes the Topic Models' unsupervised learning technique to extract Aspects from text. Various forms of Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) Topic Modeling techniques were used by researchers to extract Aspects from text [42].

LSA is used in [43] to create a feature summary with a rating for each feature. Similarly, it is employed in [44] to build a domain-dependent aspect-extraction sentiment analysis framework. LDA is used in [45] to handle the aspect-extraction task.

Titov and McDonald [46] demonstrated that both Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) methods produce wide range of topics, therefore it is difficult to identify which topics represent the aspects of the target. To solve this problem, they introduced their Multi-Grain Topic Model (MG-LDA), which produces two sets of outputs. The first output is the Global Topics in the text, whereas the second output is the Local Topics, which can zoom into the text to discover Aspects of the target.

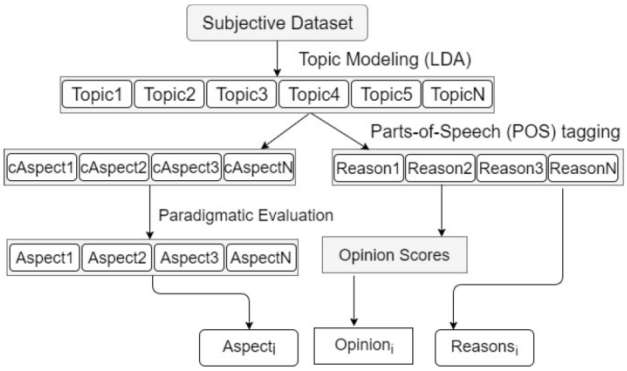

FIGURE 3. ORMFW, opinion reason mining framework [50].

Li, Huang, and Zhu [47] introduced their Dependency-Sentiment-LDA, which aims to classify sentiment of the text, then it discovers the topics inside that text. They applied their model on product reviews dataset to extract the topics of each review, show topic probabilities, and identify sentiment of each topic.

For product reviews applications, Guzman and Maalej [48] used fine grained sentiment analysis method to extract features of a target product along with their associated sentiments. They used the Natural Language Toolkit (NLTK) to extract features by finding expressions of multiple words that frequently co-occur. Then they used a lexical-based tool, SentiStrength [49], to carry out Sentiment Analysis. Finally, they used LDA Topic Model to group fine-grained aspects into more expressive high-level aspects.

ORMFW [50] is an Opinion Reason Mining Framework, which utilizes Topic Modeling to group Aspects for product reviews domain, then it links them to their reason candidates. Later, Khalid *et al.* [51] enhanced ORMFW by proposing a method that uses linguistic relations to extract implicit aspect terms and assign a weight for each term. Fig. 3 shows the ORMFW framework.

Chen *et al.* [52] developed OESTM, an On-line Evolutionary Sentiment Topic Analysis Modeling. OESTM uses a non-parametric Hierarchical Dirichlet Process (HDP) to calculate optimum number of topics for extracting aspects. They applied the model on restaurant reviews domain to track evolution of sentiment on restaurant aspects like food, taste, waiters, etc. Because OESTM uses a time-dependent Chinese Restaurant Franchise Process (CRFP) to track the evolution of topics, it faces the limitation of selecting the right time span for the used CRFP [52].

6) TRANSFER LEARNING

Main purpose of Transfer Learning is to enhance performance of learning on target domains through transfer of knowledge in divergent but related source domains [53]. Fig. 4 shows a visual example to simplify the concept of Transfer Learning.

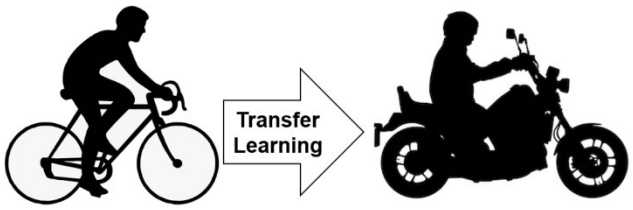

FIGURE 4. Perceptive example about transfer learning.

A person who is familiar with bicycles can transfer this knowledge to the motorcycle's domain.

Tao and Fang [54] used the relatively new Transfer Learning model BERT (Bidirectional Encoder Representations from Transformers) and XLNet, which is an autoregressive pre-training method that supports learning bidirectional contexts. Their research results show that XLNet outperforms both BERT and Deep Learning methods in most of studied cases where multi-label sentiment analysis task is carried out.

BERT always outperformed Deep Learning models, though it has the limitation of dealing with sentences of maximum length. In general, Transfer Learning models need special computing resources, otherwise the training process would be too slow.

B. SUPERVISED LEARNING

As indicated in [6], it is extremely difficult to employ Supervised Machine Learning for the Sentiment Reasoning task because of lack of labeled data that identifies majority of reasons for all sentiments. However, some studies managed to apply Supervised Learning on specific applications and domains, where it is feasible to quantify possible reasons behind an author's stance.

Kim and Hovy [55] applied Supervised Learning on products review blogs by using online review websites with user-generated pros and cons. They trained a Maximum Entropy model on product reviews using available data at both epinions.com and complaints.com. Finally, they used the framework to automatically label pros and cons of unlabeled product review blogs, and it could achieve an average accuracy of 68%. However, their approach did not sub-categorize predicted reasons into fine-grained reason categories.

Lin *et al.* [56] used Support Vector Machines (SVM) and Naive Bayes (NB) to learn perspective of an opinionated text on both document-level and sentence-level basis. To achieve this task, they created a corpus with label perspective on a document-level, however they could not label it on sentence-level, and this was a main challenge to their work. To reduce the impact of absence of sentence-level labels, they presented a Latent Sentence Perspective Model (LSPM) to recognize how perspectives are revealed inside a document.

Zaidan *et al.* [57] utilized a discriminative SVM to extract the ''rationale'' which is the text's part that supports writer's sentiment on document-level for movies' review domain. Though extracted rationale may include reason behind sentiment, sometimes it indicates motive of the writer without explicitly mention the reason. For instance, if part of a text shows that a writer prefers a specific brand name, then that text would be identified as a rationale, although it does not clarify reason behind that writer's motives. This approach is enhanced further in [58] by labeling the rationales automatically, rather than using human annotation.

Persing and Ng [59] used a Bootstrapping Algorithm to identify possible cause for reported incidents in Aviation Safety Reporting System (ASRS). They manually annotated 1,333 documents to predict reasons from unlabeled documents.

Similarly, Boltuzic and Snajder [60] prepared Corpus of Online User Comments with Arguments [61], which is a manually labeled corpus to recognize possible reasons of opinions in discussion forums for a specific domain. However, they focused on categorizing reasons on post-level, and they did not address sentence-level reasons.

Inspired by [60], Hasan and Ng [62] used Supervised Learning to carry out a sentence-level classification for opinions inside online ideological debate forums. They manually annotated reasons of each expressed opinion inside posts from four domains, which are shown in Table 1. They used their Reason-Annotated Corpus along with Maximum Entropy model, Dependency-Based Feature Extraction, and Joint Learning to discover reasons of debaters' stances from unlabeled posts. Their system could achieve accuracies between 25.1% and 39.5% for different debate domains.

The Supervised Learning approach requires data annotation work, which can be achieved either automatically [55] for product reviews or manually [62] for debate forums. However, applying this approach on Reason Mining for Twitter would be impractical because it is not possible to create a Reasons-Corpus that is large enough to cover majority of various sentiment reasons.

C. TOPIC MODELING

Utilization of Topic Models for extracting Aspects from text has been addressed in subsection A-5. However, some studies used Topic Models to monitor all discussed topics in the subjective text, even when these topics are not categorized as aspects or features. Tan *et al.* [63] explained that majority of subjective tweets explicitly indicate the reason behind positive/negative opinion in the same text, therefore, extracting topics from the subjective tweets would certainly help us interpreting the sentiment polarity and levels.

Three different approaches were used to combine Topic Modeling task and Sentiment Analysis task:

1. Mix both tasks in a common Topic-Sentiment model/framework, e.g., [64].
2. Carry out each task separately for the same set of documents, e.g., [65].
3. Consider one of the tasks as a prior to the other, e.g., [66].

In general, Topic Models use statistical methods to discover topics that appear in a set of either short or long texts.

TABLE 1. Reason classes for debate forums [62].

Domain	Stance	Reason classes
ABO	*for*	[F1] Abortion is a woman's right (26%); [F2] Rape victims need it to be legal (7%); [F3] A fetus is not human (38%); [F4] Mother's life in danger (5%); [F5] Unwanted babies are ill-treated by parents (8%); [F6] Birth control fails at times (3%); [F7] Abortion is not murder (3%); [F8] Mother is not healthy/financially solvent (4%); [F9] Others (6%)
	against	[A1] Put baby up for adoption (9%); [A2] Abortion kills a life (29%); [A3] An unborn baby is a human and has the right to live (40%); [A4] Be willing to have the baby if you have sex (14%); [A5] Abortion is harmful for women (5%); [A6] Others (3%)
GAY	*for*	[F1] Gay marriage is like any other marriage (14%); [F2] Gay people should have the same rights as straight people (36%); [F3] Gay parents can adopt and ensure a happy life for a baby (10%); [F4] People are born gay (18%); [F5] Religion should not be used against gay rights (11%); [F6] Others (11%)
	against	[A1] Religion does not permit gay marriages (18%); [A2] Gay marriages are not normal/against nature (39%); [A3] Gay parents can not raise kids properly (11%); [A4] Gay people have problems and create social issues (16%); [A5] Others (16%)
OBA	*for*	[F1] Fixed the economy (21%); [F2] Ending the wars (7%); [F3] Better than the republican candidates (25%); [F4] Makes good decisions/policies (8%); [F5] Has qualities of a good leader (14%); [F6] Ensured better healthcare (8%); [F7] Executed effective foreign policies (6%); [F8] Created more jobs (4%); [F9] Others (7%)
	against	[A1] Destroyed our economy (26%); [A2] Wars are still on (11%); [A3] Unemployment rate is high (5%); [A4] Healthcare bill is a failure(9%); [A5] Poor decision-maker (7%); [A6] We have better republicans than Obama (5%); [A7] Not eligible as a leader (20%); [A8] Ineffective foreign policies (4%); [A9] Others (13%)
MAR	*for*	[F1] Not addictive (23%); [F2] Used as a medicine (11%); [F3] Legalized marijuana can be controlled and regulated by the government (33%); [F4] Prohibition violates human rights (15%); [F5] Does not cause any damage to our bodies (6%); [F6] Others (12%)
	against	[A1] Damages our bodies (23%); [A2] Responsible for brain damage (22%); [A3] If legalized, people will use marijuana and other drugs more (12%); [A4] Causes crime (9%); [A5] Highly addictive (17%); [A6] Others (17%)

Some of these models are non-probabilistic, like the Latent Semantic Indexing (LSI) - also known as Latent Semantic Analysis (LSA) [67] and the Non-Negative Matrix Factorization (NMF) which became popular when Lee and Seung [68] employed it for Topic Modeling.

The second category of Topic Models that gained popularity during the last two decades is the probabilistic topic modeling. Hofmann [69] transformed the LSI method into a probabilistic model called Probabilistic Latent Semantic Analysis (PLSA) or Probabilistic Latent Semantic Indexing (PLSI). Later, Blei *et al.* [41] used the Latent Dirichlet Allocation (LDA) in the field of Machine Learning, then it gradually evolved to become one of the most-popular probabilistic models nowadays because of its coherent outputs, though it is slower than both PLSA and NMF models. Both Probabilistic and Non-Probabilistic categories of topic models rely on Unsupervised Learning techniques, however some Supervised versions of these models were developed to handle certain tasks like predicting response values for new set of texts [70].

When LSA and LDA methods are compared, LDA can extract more coherent topics [71]. Smatana *et al.* [72] also compared LSA and LDA by applying them on Reuter dataset and they reached to a similar conclusion. Fig. 5 demonstrates that – regardless of the selected number of topics - the coherence scores of LDA are higher than the Latent Semantic Indexing (LSI), which is the developed algorithm for LSA.

There is a need for better methods to determine LDA parameters for achieving good results, especially the ''number of topics'' parameter [73]. Furthermore, it is likely that

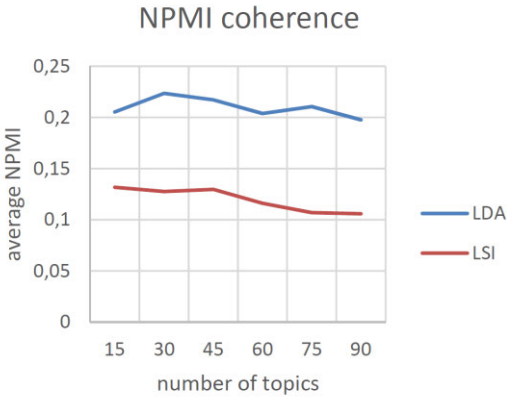

FIGURE 5. LDA vs LSI Topic coherence for Reuters dataset [72].

the wording of the extracted aspects by Topic Models would be different from the chosen wording for the manually labeled training datasets, as a result Topic Models may fail to extract some aspects due to such wording discrepancies.

Multiple forms of Topic Models were developed to address some of the challenges that face probabilistic topic modeling methods, like selecting number of topics, tuning model's parameters, identifying global topics and local/sub-topics, etc. Researchers can use these models in their programs through multiple software packages and libraries. To give you an idea, in addition to the Gensim [74] library for Python, tomotopy [75] library provides Python programmers with a wide range of Topic Modeling algorithms, which include:

- Latent Dirichlet Allocation [41]
- Hierarchical LDA [76]

- Hierarchical Dirichlet Process [77]
- Correlated Topic Model [78]
- Pachinko Allocation [79]
- Dynamic Topic Model [80]
- Hierarchical Pachinko Allocation [81]
- Multi Grain LDA [45]
- Labeled LDA [82]
- Supervised LDA [70]
- Partially Labeled LDA [83]
- Dirichlet Multinomial Regression [84]
- Generalized Dirichlet Multinomial Regression [85]

Mei *et al.* [86] developed a Topic-Sentiment Mixture (TSM) model which can discover multiple sub-topics and their associated sentiments in a set of blog documents. To track the dynamics of topics, they used a Hidden Markov Model (HMM) to mark each word in the text with sentiment and a topic. To extract sentiment, the model utilizes a commercialized sentiment-search-engine called Opinmind. This urged Pang and Lee [9] to be somehow critical about the TSM model because it borrows someone else's solution to solve the sentiment analysis problem. However, this model is useful for analyzing microblogs where multiple topics and sentiments may exist in each document, though it has limited use for short text applications - like Twitter - where each text contains a single topic most of the time.

Wang *et al.* [12] used a CRF sequence labeling along with a Topic Model to develop a sentiment reasoning model called Aspects and Sentiment of Events from Microblog (ASEM). Their model utilizes the knowledge of an event to learn about other events. It tries to discover the most-discussed topic related to an event, then it correlates that topic with the sentiment. ASEM simultaneously extracts features and feature-specific sentiment words of events.

D. DATA VISUALIZATION

By grouping document sets based on their timestamps and comparing the Topic Model's outputs for each time slot, it is possible to track birth, evolution, and death of topics. This helps users to understand possible reasons of sentiment levels for each time slot. However, it would be much faster and easier to monitor and track topics using visualization techniques instead of relying on standard Topic Modeling output, which is usually a list of topic keywords.

By monitoring curves of topics over time during a selected period, it is possible to conclude main topics which are discussed during that period. Although this interpretation technique is not fully automatic, it is still useful to analyze topics which are discussed both inside and outside the sentiment variation period [63].

In addition to Topics Over Time curves, Word Cloud Graphs - like the one shown in Fig. 6 – can be applied on the output of Topic Models to represent density of each topic within the selected document group. In this representation method, higher-frequency topics will have larger fonts. This helps decision-makers to identify the most discussed topics

FIGURE 6. Word-cloud representation [87].

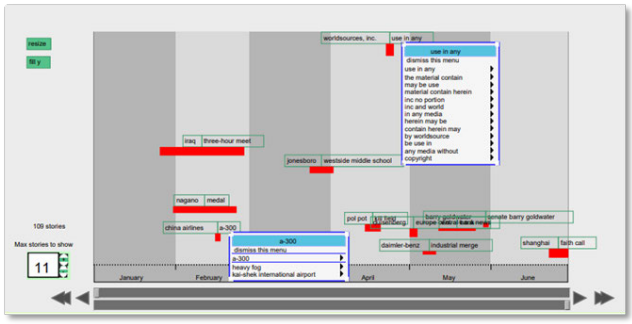

FIGURE 7. TimeMines topic visualization [88].

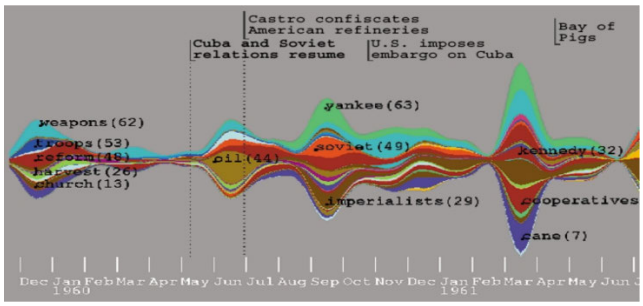

FIGURE 8. ThemeRiver topic visualization [89].

during the sentiment period, and it may lead to concluding main motives behind sentiment.

TimeMines [88] is a simple system that detects semantic features using standard statistical methods, then it groups them into topics. Finally, the system visualizes top ranked topics as shown in the example of Fig. 7. Due to simplicity of the used algorithms, the system is useful for capturing major terms only.

ThemeRiver [89] is a Topic Visualizing software that was developed to track topic evolution within a large set of documents. Fig. 8 shows a sample output of ThemeRiver which analyzes Fidel Castro's documents and interviews from year 1959 to 1961. However, as indicated by Havre *et al.* [89], the system needs a faster and a more accurate algorithm to handle user's interactive functions. For instance, zoom function is relatively slow and inaccurate.

VOSviewer [90] is a visualization software for analyzing books and articles. The software help users to identify emerging topics from a set of books or articles by viewing the bibliometric maps of the topics.

By using this software on documents which include sentiment variations, it is possible to visualize discussed topics

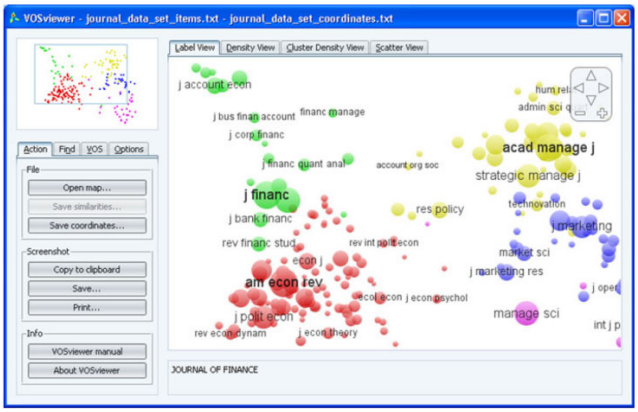

FIGURE 9. VOSviewer cluster density view of a journal map [90].

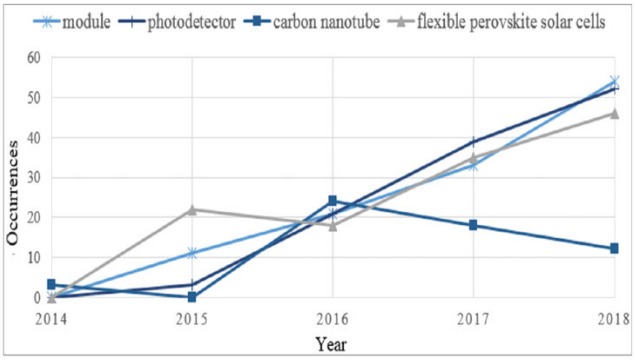

FIGURE 10. Variation in PSC term occurrence [91].

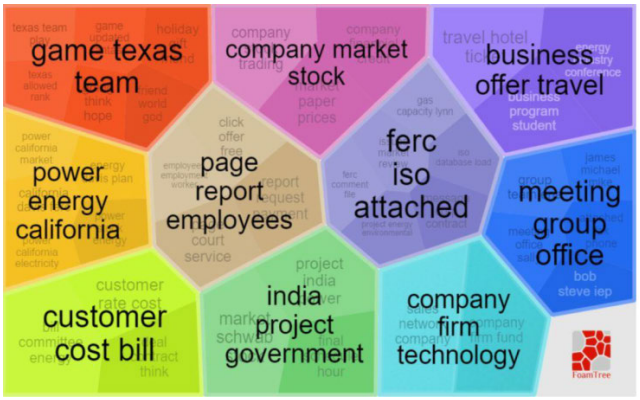

FIGURE 11. VisARTM document representation [93].

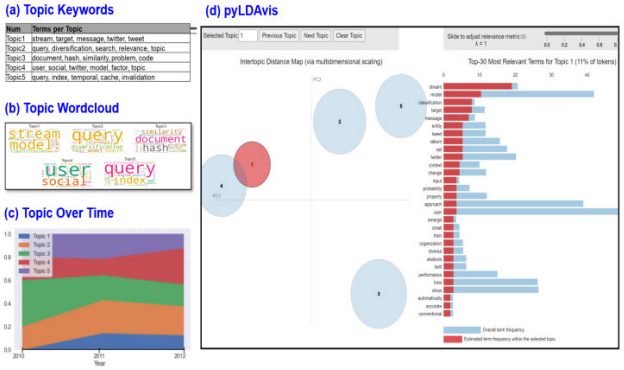

FIGURE 12. Python topic visualization for SIGIR 2010-2012 dataset.

during the selected period. Fig. 9 shows how VOSviewer assigns colors to topic terms based on their frequency and average years of publications.

With the help of VOSviewer [90] software, Shen and Wang [91] used the LOOKUP function of MS Excel and the CHART function of MS PowerPoint to capture and represent the evolution of topic terms related to their selected target, which is the Perovskite Solar Cell (PSC). Fig. 10 demonstrates the increase and decrease of attention toward research terms related to PSC.

VisARTM [92] is a web-based topic visualization tool that uses an additive regularization of Topic Modeling library called BigARTM. It represents the outputs of topic models in multiple ways.

Fedoriaka [93] used VisARTM to obtain hierarchic topic visualization using polygons which show short description of documents along with the main topic labels as shown in the Fig. 11.

Such representation ensures better understanding of each topic without the need of skimming through actual documents. However, VisARTM suffers a limitation when handling large sets of text because it is a web-based tool.

Smatana *et al.* [72] developed an interactive tool to visualize topics evolution over time. The tool can also extract sentiment of the chosen documents.

There are many other software packages that handle Topic Visualization task, however users who are familiar with Python can use its powerful libraries and tools for Topic Visualization.

To demonstrate Python's visualization abilities, LDA topic modeling is applied on a set of 294 documents which represent Information Retrieval abstracts from the proceedings of ACM SIGIR 2010-2012.

Fig. 12 shows some visualization formats that can be obtained through Python functions and codes, like the pyLDAvis [94] which was introduced for the first time in year 2015.

Topic Keywords list is a standard output of LDA topic model. However, when a simple tool like Wordcloud is used, the importance of each keyword could be easily identified.

Nonetheless, the pyLDAvis interactive tool draws the bubbles of topics in vector space. The size of each bubble represents the probability of that topic in the set of documents, whereas distances between bubbles represent similarity and possible overlaps between topics. If you click on one of the bubbles, a list of that topic's top words will be shown. The color code of the right side of the drawing reflects probability of each keyword in that topic. Finally, by drawing the probability of each topic over time, evolution of topics throughout SIGIR publication years from 2010 to 2012 could be monitored.

Dfr-browsers [95] is a free web-based tool for browsing topics from a set of documents. Fig. 13 shows sample of

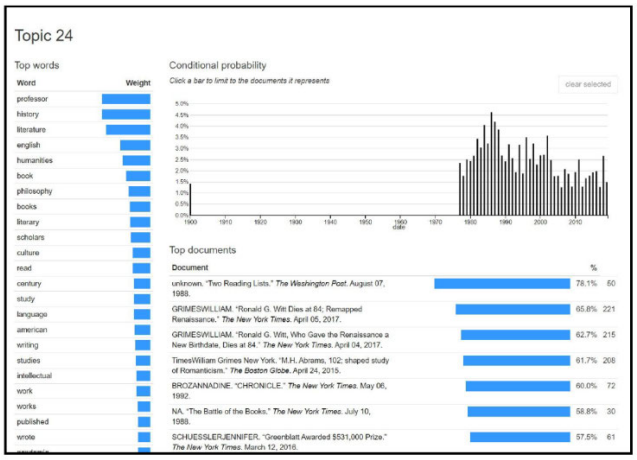

FIGURE 13. Dfr-browser topic representations [95].

dfr-browsers topic representation forms. Although most of dfr-browsers capabilities can be achieved by Python's interactive tool of pyLDAvis, dfr-browser is more user-friendly as it does not require coding skills. However, it lacks the flexibility of pyLDAvis which covers vast choices of Topic Modeling algorithms.

III. SENTIMENT VARIATIONS' REASONING

This section focuses on the three main approaches which were used to interpret public sentiment variations: (1) Event Detection method, with focus on the Topic Sentiment Change Analysis method, (2) Foreground-Background Topic Modeling approach, with focus on FB-LDA Model, and (3) Tracking Sentiment Spikes approach.

A. EVENT DETECTION

Some reason-mining studies aimed to detect and track unspecified events from social media, news, and blog sites based on the data surge these events cause in the media pipeline. In some cases, these emerging events are the main reasons behind changes in sentiments towards certain entities or targets. Therefore, detecting emerging events is a useful measure for interpreting sentiment variations over time. However, this method focuses mainly on events that cause major spikes and surges in the topics stream, hence it cannot discover lower frequency emerging topics that could be the reason for sentiment variations [63].

Outside the Sentiment Reasoning field, many event detection techniques were developed to capture spiky topics in an online data stream. Leskovec *et al.* [96] proposed a method for tracking short phrases from online text. They developed an algorithm for grouping textual variants of these phrases. Main purpose of their study was to track named entities over time.

A method was developed in [97] for detecting real-time events like earthquakes from Twitter. To detect a specific event, they used a Feature-Based classifier to group tweets. Then they applied a ''probabilistic-spatiotemporal-model'' to find location of event. By considering tweets as detected data associated with location-information, the event-detection process is simplified by detecting a target and its location-information from the received data. This is a much simpler method when compared to many ubiquitous computing methods, wherein calculating location of the target is the most important job. The system could detect 96% of Japan Meteorological Agency (JMA) earthquakes by just reading real-time tweets. Once an earthquake is detected, the system automatically sends warning emails to registered users. However, the developed system can handle one event at a time and cannot detect multiple events.

A sophisticated method for summarizing real-time event-related tweets was developed in [98]. The method uses Hidden Markov Models (HMM) to learn event's basic hidden-state representation. The event summarization process consists of two elements: (i) event-detection or event-segmentation and (ii) event summarization. HMM was used to segment events because of its ability to automatically learn variations in language models of sub-events. The developed model showed good results in summarizing events like American Football games however it was not tested for important but unpredicted events like national disasters. Moreover, it did not propose efficient measures for handling irrelevant tweets and noisy data.

Weng and Lee [99] targeted Twitter where new events are tweeted and discussed. Twitter is a challenging target for event-detection tasks as it is very scalable, and it is full of noisy data or tweets which are not related to any new event. In their study, they developed a method called EDCoW ''Event-Detection with Clustering of Wavelet-based Signals'', which uses Wavelet Transform for filtering noisy Twitter data or trivial-words. EDCoW showed good performance, however the experiment was applied to a relatively small dataset.

The above-mentioned studies focused on the Event Detection task itself, however they did not propose any mechanism for correlating events and sentiment variations. To the best of our knowledge, the first research work that correlated events with sentiment variations was the work of Jiang *et al.* [100] who tracked sudden increases in number of documents for discussed topics and correlated these spikes with the changes in overall sentiment levels. Inspired by the Topic-Sentiment Mixture (TSM) models [86], they introduced their Topic Sentiment Change Analysis (TSCA) framework. Fig. 14 is a representation of their framework which (i) aggregates sentiment levels to monitor sentiment over time, (ii) uses a rule-based method to classify sentiment, (iii) applies a time-partition technique to detect topics' spikes and identify the events that may cause sentiment changes, and (iv) evaluates the ranking of possible events that caused the change of overall sentiment level.

To carry out the topic discovery task for long multiple-topic blog articles, Jiang, *et al.* [100] used the PLSA topic model, whereas they used a rule-based classifier to detect the sentiment. By monitoring sentiment variations, they noticed that a rise of a topic popularity causes sentiment change for

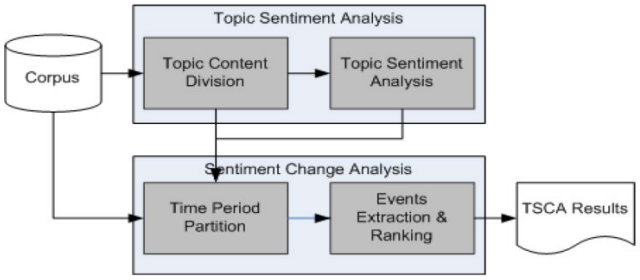

FIGURE 14. Topic sentiment change analysis framework [100].

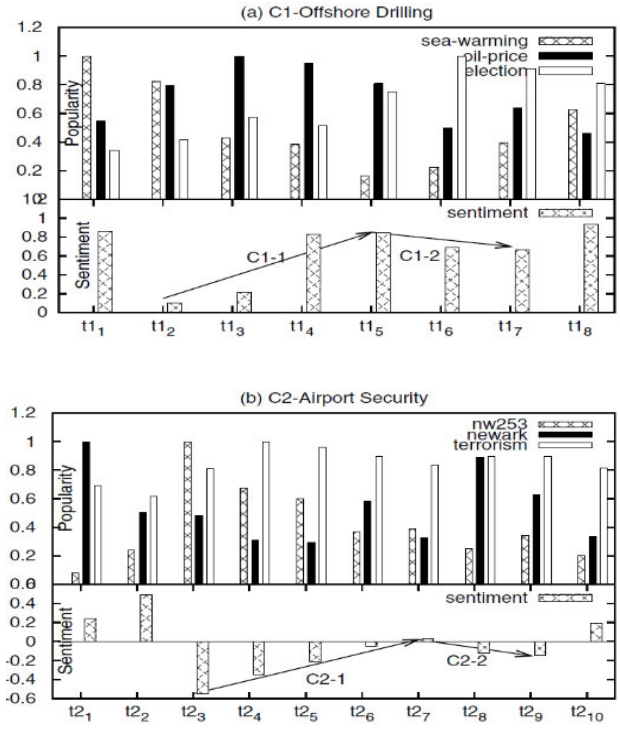

FIGURE 15. Sentiment variation reasoning using PLSA and TTP [100].

that topic. Therefore, they assumed that a sudden increase in a topic's number of documents indicates a major sentiment variation. As a result, instead of using equal partitions for the documents time slots, they used a Time Period Partition (TPP) algorithm that selects the beginning and the end of the time slot based on topics' number of documents. Once the boundaries of the time slot are decided, they identify the topic/event which has the highest probability in that time slot. Next step is calculating the total sentiment value for the sentences that fall within the selected time slot. Finally, they select the candidate topic that is adjacent to the time slot of a sentiment change.

Jiang, *et al.* [100] applied their method on two corpuses of long blog articles: Corpus (C1) which has a set of documents on the subject of ''offshore drilling'', and Corpus (C2) on the subject of ''airport security''. Fig. 15 shows the dynamics of discovered topics over time, compared with changes in sentiment over time. Given that their method detects sentiment variations based on the number of topic's documents, it may mistakenly identify a fake sentiment increase/decrease when popularity of a topic is decreased. To illustrate, the second sentiment variation C1-2 in plot (a) of Fig. 15 shows a major decrease of positive sentiment level just because of the decline of the topic of ''oil price''. This is the reason why it was decided to ignore the C1-2 sentiment variation [100].

Many other Event Detection techniques were introduced later by various researchers; however, these studies focused on enhancing the Event Detection part, without addressing the relation between sentiment levels and detected events. In [101], a method was developed for summarizing opinions on Twitter's entities through analyzing Twitter's hashtags to detect the presence of target entity then conclude polarity of identified tweet. Although utilizing the hashtag information of Tweets can be useful, relying on hashtags is a genuine limitation because the analysis excludes tweets that do not have hashtags.

Eventweet [102] is a scoring scheme for events. It tracks localized-events from live Twitter stream using a time-sliding-window technique. During the required time window, the system detects high frequency words from the live stream. Zhou *et al.* [103] used Latent Event & Category Model (LECM) to build their unsupervised framework for detecting events from Twitter. The framework creates a lexicon from online news during the same period of tweets. Then processed tweets are filtered based on their similarity to the extracted news words. As a result, tweets related to hot events/news are extracted.

A Supervised Learning approach was used in [104] by employing Neural Networks to detect events from tweets which are converted to vectors using Global Vectors for Word Representations (GloVe) embeddings. However, Hettiarachchi *et al.* [105] explained that using supervised learning methods to handle dynamic real time Twitter stream is not very efficient because of possible major discrepancies between real-life data and training data.

Peng *et al.* [105] introduced Emerging Topic detection framework based on Emerging Pattern Mining (ET-EPM). This framework transforms the standard emerging event detection into an emerging pattern clustering by using High Utility Itemset Mining (HUIM) algorithm, then they used Local Weighted Linear Regression (LWLR) for highlighting the rank of emerging topic words.

Embed2Detect [106] is a multi-language event detection system that utilizes Skip-gram word embeddings and hierarchical clustering. The method considers semantics in addition to the syntax and statistics of the text and had achieved good results in both politics and sports domains. It was also tried in [106] to use other word embedding methods like BERT, however it was realized that such advanced methods need relatively long learning times.

B. FOREGROUND-BACKGROUND TOPIS

This method was introduced by Tan *et al.* [63] who tracked changes in positive/negative sentiment, then extracted the Emerging Topics from the Foreground period – which represents the period of a positive or negative sentiment

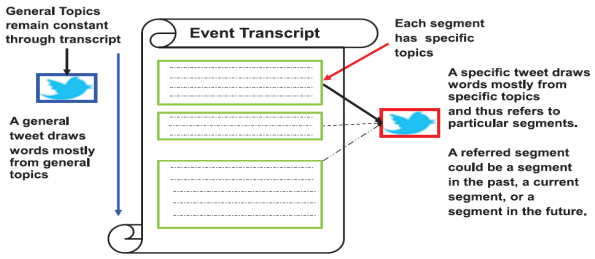

FIGURE 16. Conceptual ET-LDA model [107].

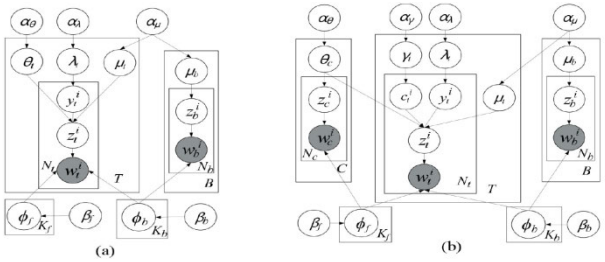

FIGURE 17. (a) FB-LDA and (b) RCB-LDA models [63].

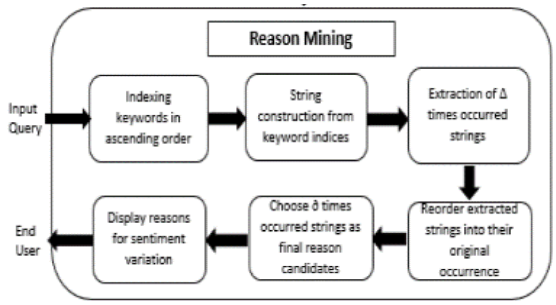

FIGURE 18. Reason mining framework using FB-LDA [110].

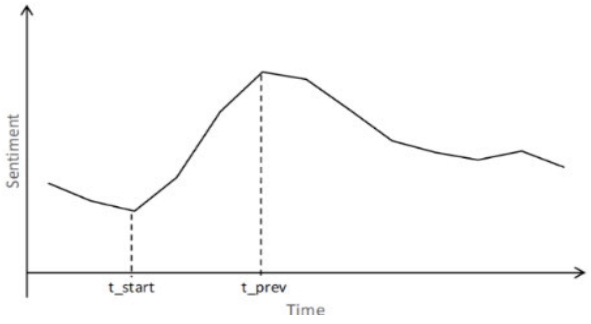

FIGURE 19. Sentiment spike [120].

variation – by removing from the Foreground all old topics that appeared in the Background period, which represents the duration before the Foreground period. For extracting the tweets' sentiment, they combined both SentiStrength and TwitterSentiment tools in a hybrid approach.

The work of Tan *et al.* [63] was inspired by the article of Hu *et al.* [107] who built an event detection and segmentation system. For an event that appears inside a large-scale group of tweets, Hu *et al.* [107] explained that the main research issues which face researchers in the event-detection and text processing fields are (i) extracting the topics which are contained in the event's text and the tweet and (ii) segmenting the event. Hu *et al.* [107] aimed to address both issues together as they realized that they are both ''inter-dependent''. They presented a Bayesian-Model called ''Event & Tweets Latent Dirchlet Allocation'', ET-LDA, which handles both tasks of modeling the topics and segmenting the events as demonstrated in Fig. 16.

Tan *et al.* [63] noticed that the emerging topics within the sentiment variation period are associated with the reasons behind sentiment variations. They employed the Foreground and Background Latent Dirichlet Allocation (FB-LDA) Model to filter foreground-topics and remove background-topics during the variation period. They assumed that emerging topics represent main reasons behind sentiment variations. Tan *et al.* [63] also used a Reason Candidate and Background LDA (RCB-LDA) model to assign ranking of topics based on their frequency/popularity. From the FB-LDA results, the RCB-LDA extracts ''representative'' tweets for the emerging topics to represent the reason candidates. Using representative tweets as reason candidates makes it easier for the user to understand the selected reasons. Then the RCB-LDA ranks reason candidates based on frequency of their category in the tweets during the variation period. Fig. 17 shows both FB-LDA and RCB-LDA models.

Although Tan *et al.* [63] did not focus on the Sentiment Analysis task itself, their Reason Mining work was followed by series of projects through other researchers – e.g. [108]–[119] - who tried to use better Sentiment Analysis classifiers for the FB-LDA Model. To illustrate, Fig. 18 shows a Reason Mining framework [110] that utilizes FB-LDA and RCB-LDA models.

C. TRACKING SENTIMENT SPIKES

This method was introduced by Giachanou, *et al* [120] who (i) extracted tweets' sentiment using SentiStrength [48] tool, (ii) detected spikes of sentiment using an outlier detection algorithm, (iii) analyzed the topics within the sentiment spike through LDA model, and (iv) ranked these topics based on their contribution to the sentiment spike using the Relative Entropy method.

In this approach, an anomaly detection method is used to detect a spike in the sentiment's trend over time [121]. This method belongs to the field of time series and it aims to discover sudden peaks in the positive or negative sentiment trend. It detects an outlier by calculating the normal residuals of each observation.

Fig. 19 shows a typical sentiment spike, where ''t_start'' represents the timestamp when the sentiment starts increasing, and ''t_prev'' represents the timestamp when the spike occurs.

Giachanou, *et al* [120] carried out LDA for the period between ''t_start'' and t_prev'' as they assumed that the topics which exit in that period have caused the spike.

Relative Entropy, also known as Kullback-Leibler divergence (KL-divergence), measures the difference between

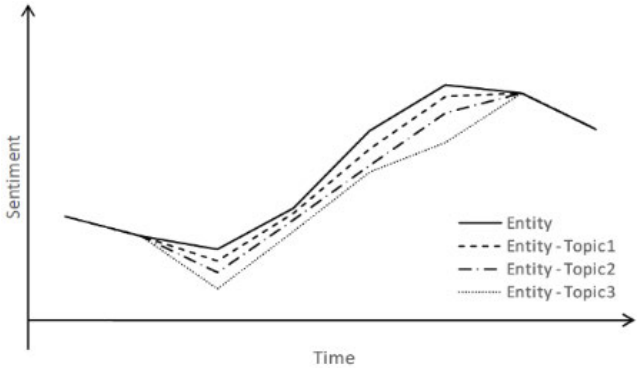

FIGURE 20. Topics inside a sentiment spike [120].

two probability distributions, and it was employed by Giachanou, *et al* [120] to rank the discovered topics inside the sentiment spike. Fig. 20 demonstrates some topic trends inside a sentiment spike.

IV. DISCUSSION

To address main research questions, which are listed in Section I, two Twitter datasets are analyzed: First one is a pre-labeled dataset in the domain of airlines services, and the second is a time-stamped unlabeled dataset in the domain of products. We manually labeled the sentiment polarity and its possible reason for the second dataset.

A. DATASETS

The first dataset is the famous US Airlines Twitter Dataset [122]. It was scrapped from Twitter in February 2015 for customer reviews of six US Airlines. Each tweet is labeled with its either negative, neutral, or positive sentiment polarity. The dataset contains 9,179 negative tweets, which were further categorized based on main reason behind sentiment, whenever that reason is explicitly mentioned the text. This dataset is used here only to address the 1st research question and to confirm our observation about Explicit and Implicit Reasons, whereas the second dataset, which we manually annotated, is used to address all research questions.

The second dataset, ''Apple Tweets'', is extracted from the Stanford Twitter Dataset, which contains 467 million tweets for a period of seven months from 01-Jun-2009 to 31-Dec-2009. The extractors of the dataset estimated that it contains 20-30% of all public tweets of the mentioned period [123]. From this dataset, all tweets about ''Apple'' for the period from 30-Jun-2009 to 03-Jul-2009 were extracted. After deleting all non-English tweets, we manually labeled the sentiment polarity of each of the remaining 7,079 tweets, which include 1,733 negative, 2,873 neutral, and 2,473 positive tweets.

B. RQ (1) EXPLICIT REASONS

By analyzing the reasons/categories of negative tweets in the US Airlines dataset – see Fig. 21 –it is noticed that 87% of the negative tweets explicitly indicate the reason behind negative

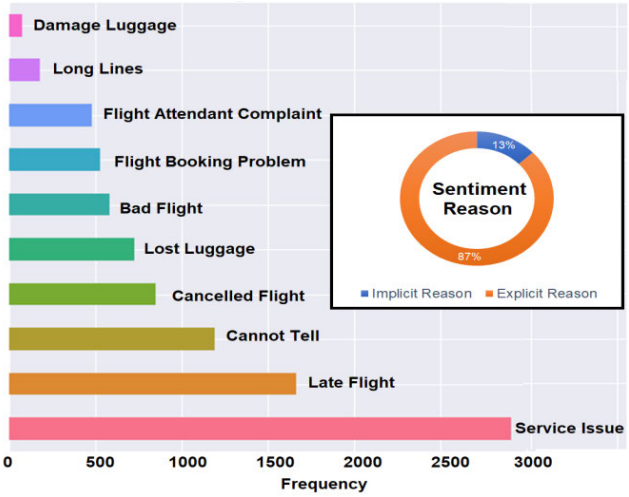

FIGURE 21. Sentiment reasons in US airlines negative tweets.

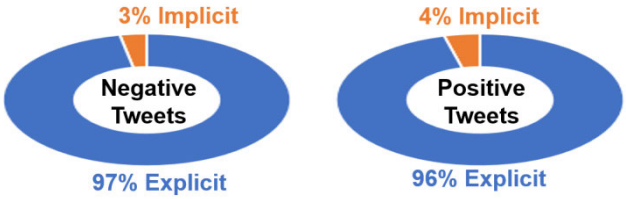

FIGURE 22. Implicit vs Explicit sentiment reasons in 'Apple'' tweets.

sentiment inside the tweet's text. However, given that the US Airlines dataset does not include reasons for positive tweets, the second dataset is used to address positive sentiment cases.

For the ''Apple Tweets'' dataset, all positive and negative tweet were analyzed, then we manually labeled the concluded reason behind sentiment. Fig. 22 shows that the reason for positive/negative sentiment is explicitly mentioned inside the text of the tweet for more than 95% of the subjective tweets.

Hence, the 1st research question is answered: most of subjective tweets explicitly indicate reasons of sentiment. Automatic detection of these explicit reasons would certainly help decision-makers to understand opinions imbedded inside tweets.

C. RQ (2) ASPECTS

To analyze sentiment variations over time, negative ''Apple'' tweets were segregated, then we aggregated the manually labeled sentiment on daily basis by counting the number of negative tweets per day. Fig. 23 shows the daily sentiment level for all Apple's 1,733 negative tweets from 30-Jun-2009 to 03-Jul-2009. The figure shows a major negative spike spreading over both 2nd and 3rd of July 2009 with around 127% increase in negative sentiment level on 02-Jul-2009.

We analyzed the labeled reason candidates and manually identified the top 6 topics which have had highest number of tweets, as shown in Table 2 . The distribution of each topic throughout the four days is shown in Fig. 24.

By manually extracting the discussed topics for each day, a genuinely good idea about the top reason candidates for the

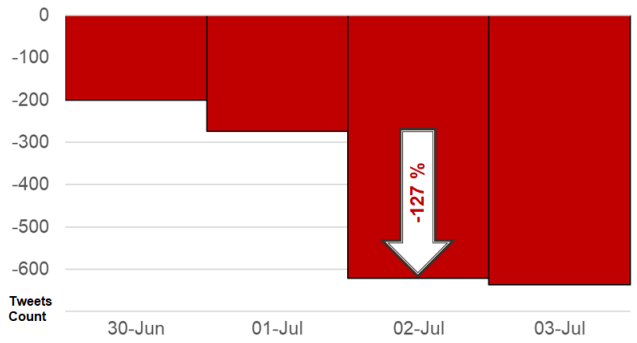

FIGURE 23. Number of Tweets inside Apple's negative sentiment spike.

TABLE 2. Reason candidates for Apple's negative sentiment.

SN.	Negative Sentiment Reason Candidate	Sample Tweet
1	Overheating	'Apple iPhone 3G S Overheated When Running CPU-Intensive Applications'
2	NVIDIA	'Apple may drop NVIDIA chips in Macs following contract fight'
3	App Rejection	'Apple Reject iKaraoke app, then files a patent for karaoke a player'
4	Child Porn	'Child porn shows up in an iPhone app, highlighting Apple's inability to regulate App Store content'
5	Vulnerability	'Not good: Apple patching serious SMS vulnerability on iPhone'
6	Store Shooting	'Woman hospitalized after shooting at Apple Store'

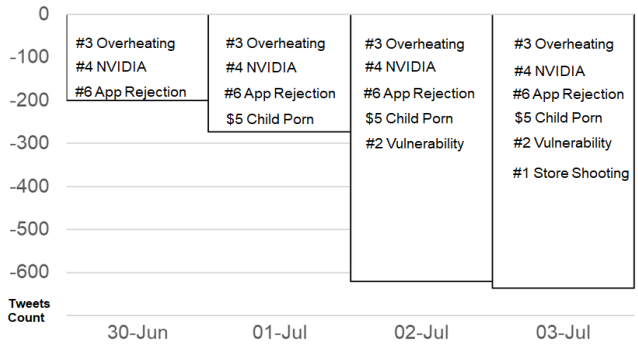

FIGURE 24. Topics Evolution inside Apple's negative sentiment spike.

expressed sentiment could be taken. In practice, Topic Modeling techniques are used to extract these reason-candidates instead of labeling them manually, and the topic models' parameters are fine-tuned to provide coherent topics which can be easily interpreted by a human.

Some of the identified reason candidates – like Overheating and Vulnerability - can be categorized as "Aspects" of our target, Apple. However, majority of these candidates are events that cannot be defined as "Aspects" or "Features" of Apple.

TABLE 3. Count of topic tweets for Apple's negative sentiment.

SN.	Reason Candidate	Tweets Count 30-Jun-2009	Tweets Count 01-Jul-2009	Tweets Count 02-Jul-2009	Tweets Count 03-Jul-2009
1	Overheating	24	27	131	92
2	NVIDIA	11	11	71	12
3	App Rejection	4	3	24	5
4	Child Porn	0	17	92	17
5	Vulnerability	0	0	144	102
6	Store Shooting	0	0	0	287

Therefore, an Aspect-Based method would not be suitable for capturing reason candidates like "Store Shooting" or "Child Porn" because it is extremely hard for any machine-learning algorithm to learn these events from a dataset, whatever large that dataset is.

Hence, the 2nd research question is now answered: Aspect-Based method cannot always capture reasons of sentiment for products and services domains.

D. RQ (3) TOPIC FREQUENCY

For "Apple Tweets", the number of tweets of each reason candidate throughout the 4 days period were counted. Table 3 shows that for 30-Jun-2009, "Overheating" has had the highest number of tweets. Therefore, identifying the highest frequency topic as the reason candidate would work well for this case. However, by analyzing the topics' counts of the next day 01-Jul-2009, you can notice that "Overheating" topic still has the highest number of tweets, however identifying it as the main reason candidate for the negative sentiment variation would be inaccurate because its number of tweets has only increased by 3 tweets from the previous day, whereas the emerging topic of "Child Porn" earned 17 new tweets on 01-Jul-2009. Therefore "Child Porn" is certainly the main reason for the jump of the negative sentiment level on 01-Jul-2009 although it does not have the highest count of tweets. Hence, this real-life Twitter example answers our 3rd research questions as follows: Main reason for a sentiment spike does not always have the highest topic frequency.

By analyzing the big sentiment level spike which spreads over both 2nd and 3rd of July 2009, you may notice that the topic which got the maximum number of tweets in this period was the "Store Shooting", although this topic did not exist at all on 02-Jul-2009 when the spike happened. This confirms our above-mentioned conclusion that the main reason for a sentiment spike does not always have the maximum count of tweets/documents. Obviously, the emerging topic "Vulnerability" is the main reason for the spike as it did not exist before 02-Jul-2009.

E. RQ (4) EVENTS

As shown in Fig. 23, there is no major change in overall "Apple Tweets" sentiment level on 01-Jul-2009, therefore Tracking Sentiment Spikes' method would not capture this

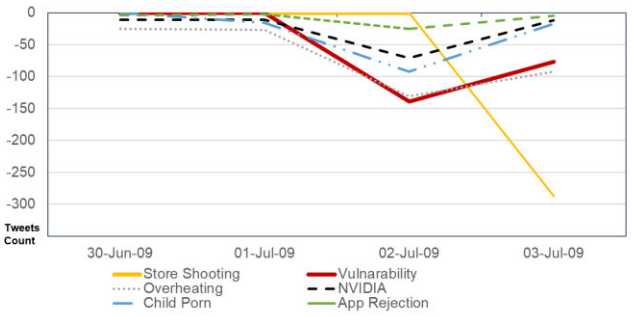

FIGURE 25. Topics Evolution inside Apple's negative sentiment spike.

small sentiment variation. As a result, the strong ''Child Porn'' reason candidate would neither be captured nor analyzed, which reveals an empirical gap in the Reason Mining method of Tracking Sentiment Spikes'. The same can be concluded for the Event Detection method as both days of 30-Jun-2009 and 01-Jul-2009 did not have major increase in any of the discussed topics. This answers our 4th research question because it indicates that the Event Detection method cannot address the Reason Mining task for the sentiment level change on 01-Jul-2009.

F. RQ (5) EMERGING TOPICS

In the above sub-sections, we concluded that the Emerging Topic of ''Overheating'' is the reason candidate on 30-Jun-2009, the Emerging Topic of ''Child Porn'' is the reason candidate on 01-Jul-2009, the Emerging Topic of ''Vulnerability'' is the reason candidate on 02-Jul-2009, and the Emerging Topic of ''Store Shooting'' is the reason candidate on 03-Jul-2009. This proves that detecting the highest frequency Emerging Topic is an efficient measure for concluding the main reason for a major sentiment variation. Therefore, it makes a lot of sense to employ the Foreground-Background Topics method for the Reason Mining task, and this answers our 5th research question.

G. RQ (6) TOPIC VISUALIZATION

To examine the role of Topic Visualization in the Sentiment Reasoning task, Fig. 25 is drawn to show the count of ''Apple Tweets'' for each topic over time. The figure clearly demonstrates that the ''Overheating'' topic was dominant during the first two days before the emergence of the ''Vulnerability'' topic on 02-Jul-2009. Although the ''Vulnerability'' topic significantly declined on 03-Jul-2009, the overall negative sentiment level was maintained because of the emergence of the ''Store Shooting'' topic.

It is therefore evident that Topic Visualization does genuinely enhance our understanding of the Sentiment Reasoning, and this answers the 6th research question.

H. RQ (7) SENTIMENT SPIKES

For ''Apple Tweets'', Fig. 26 is drawn to show the trends of ''Overall Negative Sentiment'', the ''Vulnerability'' Topic, and the ''Store Shooting'' Topic. This figure represents

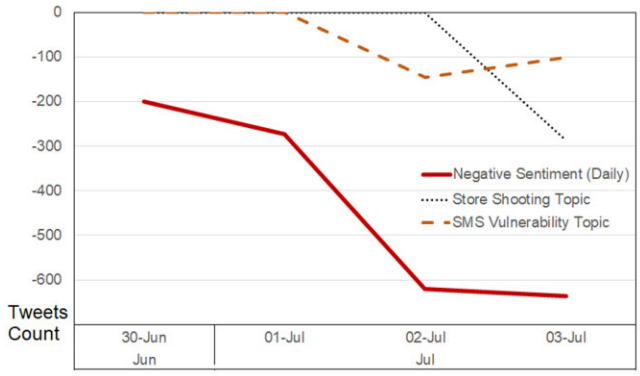

FIGURE 26. Overall Negative Sentiment vs Top Reason Candidates.

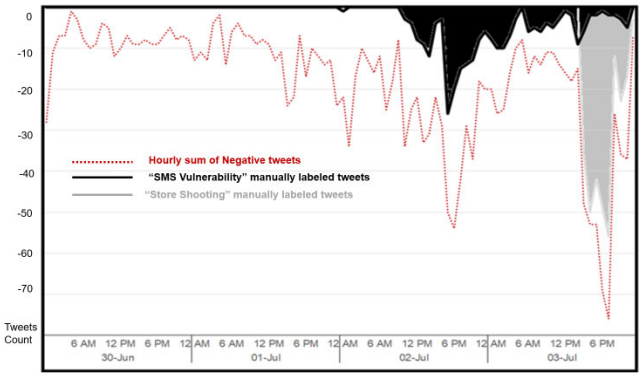

FIGURE 27. Negative Sentiment Level vs Manually Labeled Topics.

a real-life example where a major negative sentiment reason – on 3rd of July 2009 - did not cause any additional spike in overall sentiment level. The main reason for this steady-state sentiment level is the decline of some other old/background topics. Hence, the 7th research question is now answered: Sentiment variation reason does not always cause a spike in the overall sentiment level.

I. RQ (8) TOPIC MODELS

To monitor both Sentiment Trends and Topics Trends in details, the counts of manually labeled ''Vulnerability'' and ''Store Shooting'' tweets were aggregated on hourly basis as shown in Fig. 27. This visualization indicates the correlation between overall sentiment level and the trends of both topics. On 02-Jul-2009, the overall Sentiment Level was clearly controlled by the trend of the Vulnerability topic, whereas on 03-Jul-2009 the ''Store Shooting'' took control.

To address the 8th research question, instead of using our manually labeled topics or reason candidates, a practical LDA Topic Model is used to check if similar results could be obtained by employing Topic Modeling methods. MALLET [124] package is used in our experiment with Python wrapper to apply LDA with Gibbs Sampling on our ''Apple Tweets'' dataset. Our experiment filtered the topics numbers which represent the ''SMS Vulnerability'' and ''Store Shooting'' topics along with the count of documents for each topic. Fig. 28 shows remarkably similar trends

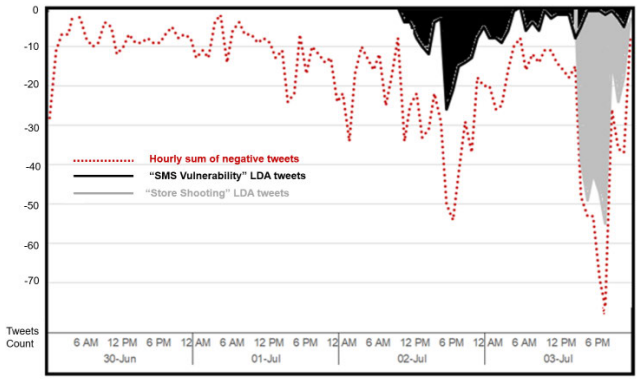

FIGURE 28. Negative Sentiment Level vs LDA Topics.

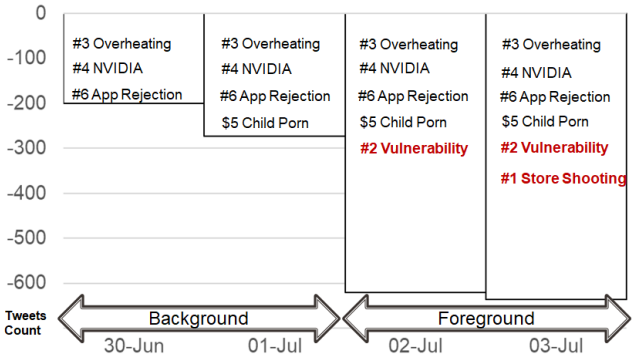

FIGURE 29. FB-LDA foreground and background periods.

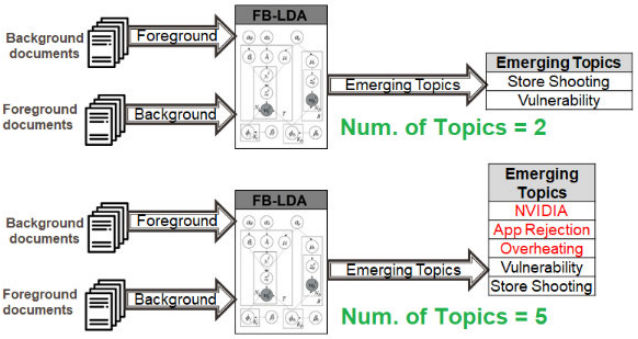

FIGURE 30. Impact of number of Topics' setting on FB-LDA.

compared to the trends of the manually labeled topics. Hence Topic Modeling and Topic Visualization techniques can provide reasonably efficient results for interpreting the reasons of sentiment levels. However, human judgement would be necessary for correlating the outputs of Topic Models and the Trend of Sentiment Level.

In general, when applying Topic Models, we identified the following main challenges that should be carefully tackled to obtain best possible results:

1. Number of Topics: Except for Hierarchical Topic Models, there is a need for identifying the number of topics' setpoint in advance. The accuracy and coherence of the model's outputs rely on the setting of this setpoint.
2. Topic Coherence: Common practice is selecting the Number of Topics that gives highest Coherence Scores. However, low number of topics may merge similar topics together, even when Coherence Scores are high. This may cause misjudgment for human analysis. Furthermore, merging similar topics would make it difficult to identify Emerging Topics correctly as these may merge with old topics.
3. High Number of Topics: To avoid manual selection of number of topics, Hierarchical Dirichlet Process (HDP) can be used. However, HDP tends to give relatively high number of topics as it identifies main topics and sub-topics. Such high number of topics needs extra human effort for analysis. Moreover, these sub-topics would be mistakenly identified as Emerging Topics if they appear during the Foreground period.

J. RQ (9) FOREGROUND-BACKGROUND TOPICS

For an automatic correlation between topics and sentiment level, we concluded that Emerging Topics Detection is the most efficient method for interpreting public sentiment variations. Therefore, to answer our 9th research question, Tan *et al.* [63] FB-LDA Model is applied on the same ''Apple Tweets'' dataset. First, our dataset is segregated into two groups of tweets. The first group represents the Foreground period – i.e., the sentiment variation period - from 2nd to 3rd of July 2009. The second group represents the Background period which consists of all tweets that appeared on 30-Jun-2009 and 01-Jul-2009. Fig. 29 demonstrates the Foreground-Background split.

As there is no clear guideline by Tan *et al.* [63] for setting the number of topics for the FB-LDA Model, we tried both settings of 2 topics and 5 topics. Fig. 30 shows that the FB-LDA model successfully detected the two emerging topics when the Number of Topics setpoint is 2, however the model added three old/background topics when the setpoint is increased to 5 topics. This shows that the FB-LDA Model is efficient in extracting the reason candidates for sentiment variations, provided that the right number of emerging topics is selected. Nevertheless, we noticed that sometimes the Topic Keywords of FB-LDA Emerging Topics include few words from the Background/old topics. For instance, the word ''Overheat'' was listed among the Topic Keywords of the ''Store Shooting'' topic. Tan *et al.* [63] had also introduced the RCB-LDA Topic Model which ranks the reason candidates. RCB-LDA should reduce the negative impact of wrong setting of Number of Topics for the FB-LDA model because it provides the rank of each emerging topic based on its contribution to the sentiment level. However, the accuracy of the ranking model still relies on the setpoint of the Number of topics because it may assign high rank for low popularity emerging topics if the model merges them with other topics due to low value of this setpoint.

Tan *et al.* [63] used a simple method to detect sentiment variations. They tracked the result value of (POS/NEG) and (NEG/POS), where POS is the sum of positive tweets,

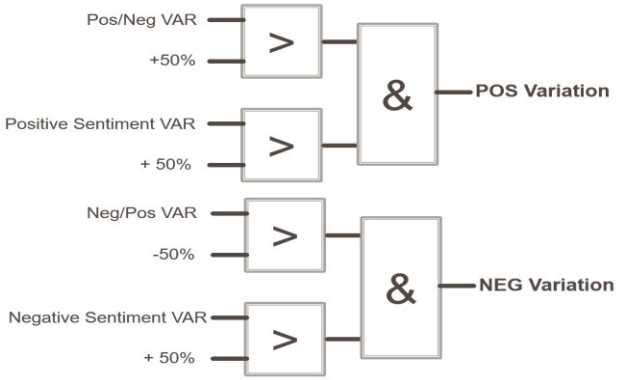

FIGURE 31. Impact of number of Topics' setting on FB-LDA.

and NEG is the sum of Negative Tweets. Whenever the result value increases by 50%, they assume a major negative/positive sentiment variation. The main advantage of this calculation method is avoiding possible misleading indications of sentiment spikes when high numbers of documents are detected just because of a sampling problem from the data source, or an impact of certain occasions like weekend periods, when some people are more likely to use Twitter. However, this method suffers from few shortcomings:

1. It identifies false major variations when quantities of both Positive and Negative tweets are too small.
2. It fails to identify major Positive and Negative variations in case both Positive and Negative events occur during the same period.
3. It identifies false Positive variation identification in case number of Negative tweets declines or reduced without any increase in Positive tweets.

Therefore, to avoid the above-mentioned limitations, we recommend that the (POS/NEG) calculation should be combined with the condition of a major increase in the Positive Sentiment Level. The same is applicable for the (NEG/POS) calculation as shown in Fig. 31.

V. CONCLUSION

Reason Mining methods help decision-makers to interpret public sentiment levels and their changes over time. Our experiments used two different real-life Twitter datasets to prove that most of subjective tweets explicitly mention the reason for positive/negative sentiment, therefore extracting the topics of the tweets is a useful measure for interpreting sentiment levels.

To extract the reasons of a certain sentiment, Aspect-Based methods provide useful outputs in the domains of products and services. However, we manually annotated a real-life Twitter dataset to demonstrate that Aspect-Based methods are not efficient when reasons of sentiment are events, even in products domain. Our experiments also showed how Topic Modeling and Data Visualization methods are helpful for carrying out the Reason Mining task. However, both methods require human judgement when main reason candidates are not the highest frequency topics.

For the Sentiment Variation Reasoning task, our experiments demonstrated that the Foreground-Background Emerging Topic Detection method is an efficient approach for interpreting public sentiment variations. We also spotted a research gap for both Event Detection method and Tracking Sentiment Spikes method as shown in real-life examples from Twitter where major sentiment reasons sometimes neither create Topic Spikes nor cause Sentiment Spikes. Furthermore, by applying FB-LDA Model on a real-life example, we could obtain good results when the number of topics' setpoint is selected correctly, however the authors of the FB-LDA model did not articulate clear guidelines for setting the number of topics. Moreover, the keywords of the FB-LDA Emerging Topics sometimes include words related to Background/old topics. Finally, we proposed an enhanced method for detecting Twitter sentiment variations to avoid the shortcomings of existing FB-LDA sentiment variations' detection method.

In our future study, a novel Reason Mining framework will be introduced to deal with the identified limitations of existing Sentiment Variations' Reasoning methods.

REFERENCES

[1] A. D'Andrea, F. Ferri, P. Grifoni, and T. Guzzo, "Approaches, tools and applications for sentiment analysis implementation," *Int. J. Comput. Appl.*, vol. 125, no. 3, pp. 26–33, Sep. 2015.
[2] Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," 2002, *arXiv:cs/0205070*. [Online]. Available: https://arxiv.org/abs/cs/0205070
[3] C. Banea, R. Mihalcea, and J. Wiebe, "A bootstrapping method for building subjectivity lexicons for languages with scarce resources," in *Proc. LREC*, 2008, pp. 2764–2767.
[4] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis," *Comput. Linguistics*, vol. 35, no. 3, pp. 399–433, Sep. 2009.
[5] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intell. Syst.*, vol. 28, no. 2, pp. 15–21, Mar. 2013.
[6] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research," *IEEE Trans. Affect. Comput.*, early access, Nov. 16, 2020, doi: 10.1109/TAFFC.2020.3038167.
[7] D. A. Miles, "A taxonomy of research gaps: Identifying and defining the seven research gaps," in *Proc. Doctoral Student Workshop, Finding Res. Gaps-Res. Methods Strategies*, Dallas, TX, USA, Aug. 2017, pp. 1–10.
[8] C. Muller-Bloch and J. Kranz, "A framework for rigorously identifying research gaps in qualitative literature reviews," in *Proc. 36th Int. Conf. Inf. Syst. (ICIS)*, Fort Worth, TX, USA, Dec. 2015, pp. 1–19.
[9] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, nos. 1–2, pp. 1–135, 2008.
[10] B. Liu, *Sentiment Analysis and Subjectivity*. London, U.K.: Chapman & Hall, 2010, pp. 627–666.
[11] L. Zhang and B. Liu, "Aspect and entity extraction for opinion mining, data mining and knowledge discovery for big data," *Stud. Big Data*, vol. 1, no. 5, pp. 1–40, 2014, doi: 10.1007/978-3-642-40837-3_1.
[12] R. Wang, W. Huang, W. Chen, T. Wang, and K. Lei, "ASEM: Mining aspects and sentiment of events from microblog," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2015, pp. 1923–1926.
[13] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining KDD*, 2004, pp. 168–177.
[14] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *Proc. Conf. Human Lang. Technol. Empirical Methods Natural Lang. Process. HLT*, 2005, pp. 9–28.
[15] C. Scaffidi, K. Bierhoff, E. Chang, M. Felker, H. Ng, and C. Jin, "Red opal: Product-feature scoring from reviews," in *Proc. 8th ACM Conf. Electron. Commerce EC*, 2007, pp. 182–191.

[16] S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. A. Rei, and J. Reynar, "Building a sentiment summarizer for local service reviews," in *Proc. WWW Workshop, NLP Inf. Explosion Era*, Beijing, China, Apr. 2008.
[17] B. O'Connor, R. Balasubramanyan, B. Routledge, and N. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in *Proc. Int. AAAI Conf. Web Social Media*, 2010, pp. 122–129.
[18] S. Moghaddam and M. Ester, "Opinion digger: An unsupervised opinion miner from unstructured product reviews," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage. - CIKM*, 2010, pp. 1825–1828.
[19] A. Ishaq, S. Asghar, and S. A. Gillani, "Aspect-based sentiment analysis using a hybridized approach based on CNN and GA," *IEEE Access*, vol. 8, pp. 135499–135512, 2020.
[20] N. Ahamed Kabeer, K. Gan, and E. Haris, "Domain-specific aspect-sentiment pair extraction using rules and compound noun lexicon for customer reviews," *Informatics*, vol. 5, no. 4, p. 45, Nov. 2018.
[21] L. Zhuang, F. Jing, and X.-Y. Zhu, "Movie review mining and summarization," in *Proc. 15th ACM Int. Conf. Inf. Knowl. Manage. - CIKM*, 2006, pp. 43–50.
[22] Y. Wu, Q. Zhang, X. Huang, and L. Wu, "Phrase dependency parsing for opinion mining," in *Proc. Conf. Empirical Methods Natural Lang. Process. EMNLP*, vol. 3, 2009, pp. 1533–1541.
[23] G. Qiu, B. Liu, J. Bu, and C. Chen, "Opinion word expansion and target extraction through double propagation," *Comput. Linguistics*, vol. 37, no. 1, pp. 9–27, Mar. 2011.
[24] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent Twitter sentiment classification," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.* Portland, OR, USA: Association for Computational Linguistic, Jun. 2011, pp. 151–160.
[25] P. More and A. Ghotkar, "A study of different approaches to aspect-based opinion mining," *Int. J. Comput. Appl.*, vol. 145, no. 6, pp. 11–15, Jul. 2016.
[26] M. Syamala and N. J. Nalini, "A deep analysis on aspect based sentiment text classification approaches," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, pp. 1795–1801, Oct. 2019.
[27] W. Jin, H. H. Ho, and R. K. Srihari, "OpinionMiner: A novel machine learning system for Web opinion mining and extraction," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining - KDD*, 2009, pp. 1195–1204.
[28] Y. Choi and C. Cardie, "Hierarchical sequential learning for extracting opinions and their attributes," in *Proc. ACL Conf. Short Papers*, 2010, pp. 269–274.
[29] P. Chen, S. Chen, and J. Liu, "Hierarchical sequence labeling model for aspect sentiment triplet extraction," in *Natural Language Processing and Chinese Computing*. Cham, Switzerland: Springer, 2020, pp. 654–666.
[30] Q. Wang and J. Ren, "Sequence prediction model for aspect-level sentiment classification," in *Proc. ECAI*, 2020, pp. 2196–2203.
[31] P. Liu, S. Joty, and H. Meng, "Fine-grained opinion mining with recurrent neural networks and word embeddings," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1433–1443.
[32] S. Jebbara and P. Cimiano, "Aspect-based sentiment analysis using a two-step neural network architecture," in *Semantic Web Challenges*. Cham, Switzerland: Springer, 2016, pp. 153–167.
[33] H. Jangid, S. Singhal, R. R. Shah, and R. Zimmermann, "Aspect-based financial sentiment analysis using deep learning," in *Proc. Companion The Web Conf. Web Conf. - WWW*, 2018, pp. 1961–1966.
[34] J. Xu, D. Chen, X. Qiu, and X. Huang, "Cached long short-term memory neural networks for document-level sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 1660–1669.
[35] S. Poria, E. Cambria, and A. Gelbukh, "Aspect extraction for opinion mining with a deep convolutional neural network," *Knowl.-Based Syst.*, vol. 108, pp. 42–49, Sep. 2016.
[36] D. D, A. Kumar Thakur, N. Prasad Diwakar, and R. V. College Of Engineering, "Aspect-based sentiment summarization with deep neural networks," *Int. J. Eng. Res.*, vol. V5, no. 5, pp. 371–375, May 2016.
[37] Q. Zhang and R. Lu, "A multi-attention network for aspect-level sentiment analysis," *Future Internet*, vol. 11, no. 7, p. 157, Jul. 2019.
[38] F. Z. R. Saraiva, T. L. C. da Silva, and J. A. F. de Macêdo, "Aspect term extraction using deep learning model with minimal feature engineering," in *Advanced Information Systems Engineering* (Lecture Notes in Computer Science), vol. 12127, V. Pant, Ed. Cham, Switzerland: Springer, 2020.
[39] B. Zohuri and M. Moghaddam, "Deep learning limitations and flaws," *Modern Approaches Mater. Sci. J.*, vol. 2, no. 3, pp. 241–250, 2020.
[40] A. K. Sharma. (2020). *Understanding Latent Dirichlet Allocation (LDA)*. Accessed: Jan. 30, 2021. [Online]. Available: https://www.mygreatlearning.com/
[41] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
[42] B. Lu, M. Ott, C. Cardie, and B. K. Tsou, "Multi-aspect sentiment analysis with topic models," in *Proc. IEEE 11th Int. Conf. Data Mining Workshop*, Dec. 2011, pp. 81–88.
[43] Y. Lu, C. Zhai, and N. Sundaresan, "Rated aspect summarization of short comments," in *Proc. 18th Int. Conf. World Wide Web - WWW*, 2009, pp. 131–140.
[44] R. O. Bueno, A. F. Bruzón, C. M. Cuza, Y. Gutiérrez, and A. Montoyo, "UO_UA: Using latent semantic analysis to build a domain-dependent sentiment resource," in *Proc. 8th Int. Workshop Semantic Eval. (SemEval)*, 2014, pp. 773–778.
[45] S. Brody and N. Elhadad, "An unsupervised aspect-sentiment model for online reviews," in *Proc. Hum. Lang. Technol., Annu. Conf. North American Chapter Assoc. Comput. Linguistics*, 2010, pp. 804–812.
[46] I. Titov and R. McDonald, "Modeling online reviews with multi-grain topic models," in *Proc. 17th Int. Conf. World Wide Web - WWW*, 2008, pp. 111–120.
[47] F. Li, M. Huang, and X. Zhu, "Sentiment analysis with global topics and local dependency," in *Proc. 24th AAAI Conf. Artif. Intell. AAAI*, vol. 24, 2010, pp. 1371–1376.
[48] E. Guzman and W. Maalej, "How do users like this feature? A fine grained sentiment analysis of app reviews," in *Proc. IEEE 22nd Int. Requirements Eng. Conf. (RE)*, Aug. 2014, pp. 153–162.
[49] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2544–2558, Dec. 2010.
[50] M. Taimoor and S. Khalid, "A novel opinion reason mining framework exploiting linguistic associations," in *Proc. 6th Int. Conf. Adv. Comput. Commun. Inf. Technol. CCIT*, Apr. 2018, pp. 6–10.
[51] S. Khalid, M. H. Aslam, and M. T. Khan, "Opinion reason mining: Implicit aspects beyond implying aspects," in *Proc. 21st Saudi Comput. Soc. Nat. Comput. Conf. (NCC)*, Apr. 2018, pp. 1–5.
[52] Y. Chen, C. Y. Yin, Y. J. Lin, and W. Zuo, "On-line evolutionary sentiment topic analysis modeling," *Int. J. Comput. Intell. Syst.*, vol. 11, pp. 634–651, Jan. 2018.
[53] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proc. IEEE*, vol. 109, no. 1, pp. 43–76, Jan. 2021.
[54] J. Tao and X. Fang, "Toward multi-label sentiment analysis: A transfer learning based approach," *J. Big Data*, vol. 7, no. 1, pp. 1–26, Dec. 2020.
[55] S.-M. Kim and E. Hovy, "Automatic identification of pro and con reasons in online reviews," in *Proc. COLING/ACL Main Conf. Poster Sessions*, 2006, pp. 483–490.
[56] W.-H. Lin, T. Wilson, J. Wiebe, and A. Hauptmann, "Which side are you on?: Identifying perspectives at the document and sentence levels," in *Proc. 10th Conf. Comput. Natural Lang. Learn. CoNLL-X*, 2006, pp. 109–116.
[57] O. Zaidan, J. Eisner, and C. Piatko, "Using annotator rationales to improve machine learning for text categorization," in *Proc. Main Conf. Hum. Lang. Technol., Conf. North Amer. Chapter Assoc. Comput. Linguistics*. Rochester, NY, USA: Association for Computational Linguistics, 2007, pp. 260–267.
[58] A. Yessenalina, Y. Choi, and C. Cardie, "Automatically generating annotator rationales to improve sentiment classification," in *Proc. ACL Conf. Short Papers*, 2010, pp. 336–341.
[59] I. Persing and V. Ng, "Semi-supervised cause identification from aviation safety reports," in *Proc. Joint Conf. 47th Annu. Meeting ACL 4th Int. Joint Conf. Natural Lang. Process. AFNLP, ACL-IJCNLP*, vol. 2, 2009, pp. 843–851.
[60] F. Boltužic and J. Šnajder, "Back up your stance: Recognizing arguments in online discussions," in *Proc. 1st Workshop Argumentation Mining*, 2014, pp. 49–58.
[61] ComArg. (2014). *Corpus of Online User Comments with Arguments*. [Online]. Available: http://takelab.fer.hr/data/comarg/
[62] K. S. Hasan and V. Ng, "Why are you taking this stance? Identifying and classifying reasons in ideological debates," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 751–762.
[63] S. Tan, Y. Li, H. Sun, Z. Guan, X. Yan, J. Bu, C. Chen, and X. He, "Interpreting the public sentiment variations on Twitter," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 5, pp. 1158–1170, May 2014.

[64] K. Eguchi and V. Lavrenko, "Sentiment retrieval using generative models," in *Proc. Conf. Empirical Methods Natural Lang. Process. - EMNLP*, 2006, pp. 345–354.
[65] F. Hurst and K. Nigam, "Retrieving topical sentiments from online document collections," in *Document Recognition and Retrieval XI*, vol. 5296. San Jose, CA, USA: SPIE, Jan. 2004.
[66] K. Eguchi and C. Shah, "Opinion retrieval experiments using generative models: Experiments for the TREC 2006 blog track," in *Proc. 15th Text REtrieval Conf. (TREC)*, Gaithersburg, MD, USA, Nov. 2006.
[67] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.
[68] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.
[69] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 1999, pp. 50–57.
[70] D. M. Blei and J. D. McAuliffe, "Supervised topic models," 2010, *arXiv:1003.0783*. [Online]. Available: https://arxiv.org/abs/1003.0783
[71] C. Chiru, T. Rebedea, and S. Ciotec, "Comparison between LSA-LDA-lexical chains," in *Proc. WEBIST*, 2014, pp. 255–262.
[72] M. Smatana, V. Martinkova, D. Maršálekova, and P. Butka, "Interactive tool for visualization of topic models," *Acta Electrotechnica et Inf.*, vol. 19, no. 2, pp. 45–50, 2019.
[73] T. Cvitanic, B. Lee, H. I. Song, K. Fu, and D. Rosen, "LDA v. LSA: A comparison of two computational text analysis tools for the functional categorization of patents," in *Proc. Int. Conf. Case-Based Reasoning*. Alexandria, VA, USA: National Science Foundation, Jan. 2016.
[74] *Gensim Project Description, Ver. 3.8.3*. Accessed: Jan. 30, 2021. [Online]. Available: https://pypi.org/project/gensim/
[75] *Tomotopy Documentation, Ver. 0.10.0*. Accessed: Jan. 30, 2021. [Online]. Available: https://pypi.org/project/tomotopy
[76] T. L. Griffiths, M. I. Jordan, J. B. Tenenbaum, and D. M. Blei, "Hierarchical topic models and the nested Chinese restaurant process," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 17–24.
[77] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Sharing clusters among related groups: Hierarchical Dirichlet processes," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1385–1392.
[78] D. Blei and J. Lafferty, "Correlated topic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 147–154.
[79] W. Li and A. McCallum, "Pachinko allocation: DAG-structured mixture models of topic correlations," in *Proc. 23rd Int. Conf. Mach. Learn. - ICML*, 2006, pp. 577–584.
[80] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proc. 23rd Int. Conf. Mach. Learn. - ICML*, 2006, pp. 113–120.
[81] D. Mimno, W. Li, and A. McCallum, "Mixtures of hierarchical topics with pachinko allocation," in *Proc. 24th Int. Conf. Mach. Learn. - ICML*, 2007, pp. 633–640.
[82] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora," in *Proc. Conf. Empirical Methods Natural Lang. Process. EMNLP*, vol. 1, 2009, pp. 248–256.
[83] D. Ramage, C. D. Manning, and S. Dumais, "Partially labeled topic models for interpretable text mining," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining - KDD*, 2011, pp. 457–465.
[84] D. Mimno and A. McCallum, "Topic models conditioned on arbitrary features with Dirichlet-multinomial regression," 2012, *arXiv:1206.3278*. [Online]. Available: http://arxiv.org/abs/1206.3278
[85] M. Lee and M. Song, "Incorporating citation impact into analysis of research trends," *Scientometrics*, vol. 124, no. 2, pp. 1191–1224, Aug. 2020.
[86] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic sentiment mixture: Modeling facets and opinions in Weblogs," in *Proc. 16th Int. Conf. World Wide Web - WWW*, 2007, pp. 171–180.
[87] M. Hofmann and A. Chisholm, *Text Mining and Visualization Case Studies Using Open-Source Tools*. New York, NY, USA: CRC Press, 2016.
[88] R. Swan and D. Jensen, "TimeMines: Constructing timelines with statistical models of word usage," in *Proc. KDD Workshop Text Mining*, 2000, pp. 73–80.
[89] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "ThemeRiver: Visualizing thematic changes in large document collections," *IEEE Trans. Vis. Comput. Graphics*, vol. 8, no. 1, pp. 9–20, Aug. 2002.
[90] N. J. van Eck and L. Waltman, "Software survey: VOSviewer, a computer program for bibliometric mapping," *Scientometrics*, vol. 84, no. 2, pp. 523–538, Aug. 2010.
[91] X. Shen and L. Wang, "Topic evolution and emerging topic analysis based on open source software," *J. Data Inf. Sci.*, vol. 5, no. 4, pp. 126–136, Sep. 2020.
[92] K. Vorontsov, O. Frei, M. Apishev, P. Romov, and M. Dudarenko, "BigARTM: Open source library for regularized multimodal topic modeling of large collections," in *Analysis of Images, Social Networks and Texts (AIST)* (Communications in Computer and Information Science (CCIS)). Cham, Switzerland: Springer, 2015, pp. 370–384.
[93] D. S. Fedoriaka, *Hierarchic Topic Models Visualization*. Dolgoprudny, Russia: MIPT, 2016.
[94] P. History. *Dates of Each pyLDAvis Version*. Accessed: Jan. 30, 2021. [Online]. Available: https://pyldavis.readthedocs.io
[95] A. Goldstone, S. Galán, C. L. Lovin, A. Mazzaschi, and L. Whitmore, "An interactive topic model of signs," *Signs J.*, Oct. 2014. Accessed: Jan. 30, 2021. [Online]. Available: http://signsat40.signsjournal.org/topic-model
[96] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining - KDD*, 2009, pp. 497–506.
[97] T. Sakaki, M. Okazaki, and Y. Matsu, "Earthquake shakes Twitter users: Real-time event detection by social sensors," in *Proc. 19th Int. Conf. World Wide Web (WWW)*, Raleigh, NC, USA, Apr. 2010, pp. 851–860.
[98] D. Chakrabarti and K. Punera, "Event summarization using tweets," in *Proc. 15th Int. Conf. Weblogs Social Media ICWSM*, Barcelona, Spain, Jul. 2011.
[99] J. Weng and B. Lee, "Event detection in Twitter," in *Proc. Int. AAAI Conf. Web Social Media*, 2011, pp. 401–408.
[100] Y. Jiang, W. Meng, and C. Yu, "Topic sentiment change analysis," in *Machine Learning and Data Mining in Pattern Recognition*. Berlin, Germany: Springer, 2011, pp. 443–457.
[101] X. Meng, F. Wei, X. Liu, M. Zhou, S. Li, and H. Wang, "Entity-centric topic-oriented opinion summarization in Twitter," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining - KDD*, 2012, pp. 379–387.
[102] H. Abdelhaq, C. Sengstock, and M. Gertz, "EvenTweet: Online localized event detection from Twitter," *Proc. VLDB Endowment*, vol. 6, no. 12, pp. 1326–1329, Aug. 2013.
[103] D. Zhou, L. Chen, and Y. He, "An unsupervised framework of exploring events on Twitter: Filtering, extraction and categorization," in *Proc. AAAI Conf. Artif. Intell.*, 2015, pp. 2468–2475.
[104] G. Chen, Q. Kong, and W. Mao, "Online event detection and tracking in social media based on neural similarity metric learning," in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Jul. 2017, pp. 182–184.
[105] M. Peng, S. Ouyang, J. Zhu, J. Huang, H. Wang, and J. Yong, "Emerging topic detection from microblog streams based on emerging pattern mining," in *Proc. IEEE 22nd Int. Conf. Comput. Supported Cooperat. Work Design ((CSCWD))*, May 2018, pp. 259–264.
[106] H. Hettiarachchi, M. Adedoyin-Olowe, J. Bhogal, and M. M. Gaber, "Embed2Detect: Temporally clustered embedded words for event detection in social media," *J. Mach. Learn.*, pp. 1–33, Sep. 2020. Accessed: Feb. 28, 2021. [Online]. Available: https://arxiv.org/pdf/2006.05908v3.pdf
[107] Y. Hu, A. John, F. Wang, and D. D. Seligmann, "ET-LDA: Joint topic modeling for aligning events and their Twitter feedback," in *Proc. 26th AAAI Conf. Artif. Intell.*, Vancouver, BC, Canada, 2012, pp. 1–7.
[108] D. Ingule and G. Chhajed, "Survey of public sentiment interpretation on Twitter," *Int. J. Eng. Res. Gen. Sci.*, vol. 2, no. 6, pp. 1–5, 2014.
[109] W. Poonam and M. Kinikar, "Interpreting the public sentiment with emotions on Twitter," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 2, no. 12, pp. 2–6, Dec. 2014.
[110] A. Patil, R. R. Sedamkar, and S. Gupta, "A novel reason mining algorithm to analyze public sentiment variations on Twitter and facebook," *Int. J. Appl. Inf. Syst.*, vol. 10, no. 3, pp. 23–29, Dec. 2015.
[111] A. Bhalerao and T. Dange, "Interpretation of public sentiment variations using tweets," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 3, no. 6, pp. 5828–5834, 2015.
[112] R. Urega and M. Devapriya, "Public sentiment interpretation on Twitter—A survey," *Int. J. Eng. And Comput. Sci.*, vol. 4, no. 8, pp. 13910–13916, 2015.
[113] R. A. Jamadar, V. Avachar, O. Bhosale, P. Khot, and P. Bramhane, "Survey on interpreting the public sentiment variations on Twitter," *J. Emerg. Technol. Innov. Res.*, vol. 3, no. 5, pp. 2349–5162, 2016.
[114] A. Kamini and B. Ezhillarasi, "Interpretation of sentiment variations on micro Blogs with wall filtering," *Int. J. Adv. Res. Trends Eng. Technol.*, vol. 2, no. 8, pp. 2394–3777, Feb. 2015.

[115] V. Signs, P. B. Avachar, B. O, P Khot, and R. A. Jamadar, "Implementation on interpreting the public sentiment variations on Twitter," *Int. Eng. Res. J.*, vol. 2, no. 3, pp. 1181–1184, 2016.
[116] S. Patil and S. Kulkarni, "Mining social media data for understanding Students' learning experiences using memetic algorithm," *Mater. Today, Proc.*, vol. 5, no. 1, pp. 693–699, 2018.
[117] R. Jeevitha, "Interpreting sentimental analysis for customer commands on E-commerce," *Int. J. Adv. Res. Biol. Eng. Sci. Technol.*, vol. 2, no. 10, pp. 928–929, 2016.
[118] R. Manikandan and A. Kalpana, "Tracking and analyzing the public opinions by interpretation techniques," *Int. J. Innov. Res. Technol.*, vol. 2, no. 11, p. 2349, 2016.
[119] P. Admane, S. Dhatrak, G. Ghanghao, V. Gawai, and M. Sonawane, "Interpreting the public sentiment variation on social media" *Global J. Adv. Eng. Technol.*, vol. 5, no. 1, pp. 68–71, 2016.
[120] A. Giachanou, I. Mele, and F. Crestani, "Explaining sentiment spikes in Twitter," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2016, pp. 2263–2268.
[121] A. Giachanou and F. Crestani, "Tracking sentiment by time series analysis," in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2016, pp. 1037–1040.
[122] Crowdflower. (2015). *Twitter US Airline Sentiment*. Accessed: Jan. 30, 2021. [Online]. Available: https://www.kaggle.com/crowdflower/twitter-airline-sentiment
[123] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proc. 19th Int. Conf. World Wide Web*. Raleigh, NC, USA, New York, NY, USA: Association for Computing Machinery, Apr. 2010, pp. 591–600.
[124] Mallet. *MAchine Learning for LanguagE Toolkit Website*. Accessed: Jan. 30, 2021. [Online]. Available: http://mallet.cs.umass.edu

FUAD ALATTAR was born in Amman, Jordan. He received the B.S. degree in electrical engineering from the University of Baghdad, in 1994, the M.S. degree in communications engineering from The University of Jordan, in 1998. He is currently pursuing the Ph.D. degree in computer science with The British University in Dubai, United Arab Emirates.

He is currently working as a Senior Vice President with Siemens in the Middle East region, managing the digital enterprise services business unit. He has more than 25 years' experience in automation engineering and digitalization fields, including positions of a General Manager, an Operations Manager, a Sales Director, and an Engineering Manager with Main Automation Contracting (MAC) companies.

He is a Designer of Automation System for the First Smart Multi-Aircraft Docking System in The World, in 1998, the Automation Systems for the Largest Water Transmission SCADA in Oman, in 2002, the Automation Systems for the Largest RO Plant of its Type in The World, in 2005, the Industrial Network for the Largest Oil and Gas Onshore-Offshore Network in the Middle East, in 2007, the Industrial Network for the Largest Irrigation SCADA in the Middle East, in 2011, and the Automation System of H2O Award Winning Project, in 2012. He has been an Active Researcher in *Signal Processing*, *Machine Learning*, *Natural Language Processing, Knowledge Management*, and *ICS Cyber Security* fields.

KHALED SHAALAN is currently a Professor of Computer Science with The British University in Dubai, UAE. He is an Honorary Fellow with the School of Informatics, The University of Edinburgh, U.K. He has an extensive experience in academic administration and management with leadership responsibilities, which include postgraduate program management, team leadership, new curriculum design and development for both undergraduate and postgraduate program, and membership of faculty and university levels committees. He founded and led the Natural Language Processing Research Group and successfully secured internal and external funds and conducted research with international collaborators. The scientific production and contribution result in high quality journal articles, communications, and presentations at conferences, dissertations, and book chapters. As more and more Arabic textual information becomes available through the Web in homes and businesses, via Internet and Intranet services, there is an urgent need for technologies and tools to process the relevant information. He is inspired by the dual goals of gaining novel insights into analysis and synthesis of written Arabic through NLP and developing Arabic NLP systems capable of providing assistance in translating Arabic text, teaching Arabic and providing feedback to learners, retrieving relevant Arabic documents, or extracting information from Arabic web contents, and Arabic Question Answering. He has published more than 250 articles and his H-Index using Google Scholar's H-index is more than 40. His research interests include topics in AI, Arabic NLP, knowledge management, health informatics, and educational technology. His major research interest includes computational linguistics in particular Arabic natural language processing (NLP).

He acts as the Chair of international conferences, journals, and amp; a Books Editor, a Keynote Speaker, and an External Member of promotions committees.