IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 3, 2021, accepted April 13, 2021, date of publication April 15, 2021, date of current version April 29, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3073657

Using Artificial Intelligence to Understand What Causes Sentiment Changes on Social Media

FUAD ALATTAR AND KHALED SHAALAN
Faculty of Engineering and IT, The British University in Dubai, Dubai 345015, United Arab Emirates
Corresponding author: Fuad Alattar (fuad.alattar@hotmail.com)

ABSTRACT Sentiment Analysis tools allow decision-makers to monitor changes of opinions on social media towards entities, events, products, solutions, and services. These tools provide dashboards for tracking positive, negative, and neutral sentiments for platforms like Twitter where millions of users express their opinions on various topics. However, so far, these tools do not automatically extract reasons for sentiment variations, and that makes it difficult to conclude necessary actions by decision-makers. In this paper, we first compare performance of various Sentiment Analysis classifiers for short texts to select the top performer. Then we present a Filtered-LDA framework that significantly outperformed existing methods of interpreting sentiment variations on Twitter. The framework utilizes cascaded LDA Models with multiple settings of hyperparameters to capture candidate reasons that cause sentiment changes. Then it applies a filter to remove tweets that discuss old topics, followed by a Topic Model with a high Coherence Score to extract Emerging Topics that are interpretable by a human. Finally, a novel Twitter's sentiment reasoning dashboard is introduced to display the most representative tweet for each candidate reason.

INDEX TERMS Emerging Topic Detection, interpreting sentiment variations, opinion reason mining, Sentiment Analysis, Sentiment Reasoning, Sentiment Spikes, Topic Model, Artificial Intelligence, Machine Learning, Filtered-LDA, FB-LDA.

I. INTRODUCTION

Hundreds of millions of tweets are being posted every day to discuss various topics [1] like politics, products, news, celebrities, etc. This rich source of users' feedbacks makes it essential for many decision-makers to persistently monitor Twitter and other social media platforms. Luckily, there are many software applications that can handle this task as illustrated in Table 1 examples. Such tools can monitor sentiment changes and spikes about specific targets, however, so far, none of the available tools has taken a step ahead by extracting possible reasons behind these sentiment variations.

Due to lack of specialized Sentiment Reasoning software applications so far, some users utilized available Topic Visualization methods to track evolution of topics and visually correlate curves of Topics Over Time with sentiment trends. For instance, Yin *et al.* [3] attempted to interpret changes of public sentiment towards Covid-19 on Twitter. They used Dynamic Topic Model (DTM) [4] to monitor evolution of topics over time, then they manually linked some of these topics to the changes of sentiments in Covid-19 tweets. However, given that there are more than 8 million tweets in the studied dataset, it is hard to verify concluded reasons as they rely on accuracy of the manually selected number of topics based on Coherence Scores of DTM. Moreover, we shall show later that highest Coherence Scores do not guarantee accurate tracking of topics over time.

Some researchers decided to tackle the challenge of interpreting public sentiment and understanding its changes over

TABLE 1. Some sentiment analysis software applications [2].

Name	Analyze "Live" Text	Free Plan	Visualization
MonkeyLearn	No	Yes	No
IBM Watson	Yes	Yes	No
Lexalytics	Yes	No	Yes
MeaningCloud	Yes	Yes	No
Rosette	Yes	No	No
Repustate	Yes	Yes	No
Clarabridge	No	No	No
Aylien	Yes	No	No
SYSTRAN.io	Yes	Unknown	No
Twinword Text Analysis Bundle	Yes	Yes	Yes

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

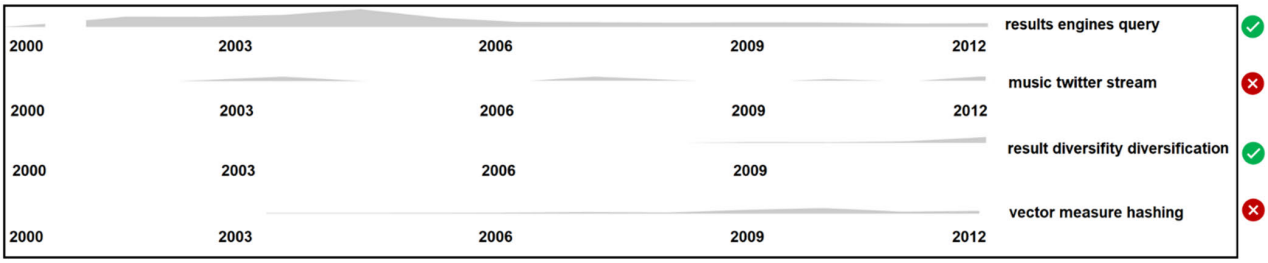

FIGURE 1. Topics over time for SIGIR dataset using dfr-browser [7].

time. Nevertheless, Poria *et al.* [5] predicted that Sentiment Reasoning will be among the top future directions of Sentiment Analysis field. In this paper, we focus on the problem of automatic discovery of reasons behind sentiment variations on social media.

A. EXISTING SENTIMENT REASONING METHODS

Sentiment Reason Mining aims to resolve two problems: first is finding the reason of a sentiment, and second is interpreting sentiment variations. Many methods were introduced to address the first problem, including Aspect-Based methods, Supervised Learning, Topic Modeling, and Data Visualization [8]. Though good research progress has been made on this branch, only few researchers decided to tackle the second problem so far. Three main approaches had handled interpreting sentiment variations. These are (1) Tracking Sentiment Spikes, (2) Foreground-Background Latent Dirichlet Allocation (FB-LDA), and (3) Event Detection. In this section, we briefly describe these approaches and identify their main limitations, which were detailed earlier in [8].

1) TRACKING SENTIMENT SPIKES

Giachanou and Crestani [9] used SentiStrength [10] tool to monitor sentiment level of tweets, then they used an outlier detection algorithm to discover sentiment spikes. Next step includes applying LDA on the tweets of the spike to identify the topic that has highest frequency as it is assumed this topic is the main reason for the sentiment spike.

Though this technique can identify reasons of sentiment variations in some cases, it is based on an inaccurate assumption that major sentiment variations always cause overall sentiment spike [8].

Moreover, this method relies on the accuracy of LDA for tracking evolution of Topics Over Time. Fig. 1 shows an example where we applied LDA on a SIGIR dataset which contains 924 abstracts from Information Retrieval subjects throughout the period from year 2000 to 2012. After we manually labeled the topics of the dataset to identify the correct number of topics ($K = 17$) which also ensured highest coherence score for the Topic Model. However, LDA failed to track evolution of ''Feature Space Hashing'' and ''Social Network Twitter'' topics as both should not appear as research trends before the year 2010. For instance, LDA trend shows Twitter topic in year 2003, though Twitter platform was created only few years later. The method of Tracking Sentiment Spikes did not consider necessary measures to avoid merging low-frequency new topics with old topics in LDA output.

2) FB-LDA

FB-LDA Model was developed by Tan *et al.* [6] who manually analyzed real-life tweets on certain targets and noticed that main reasons for sentiment variations are causally linked to Emerging Topics. They traced variations of sentiment level to identify the Foreground period when variation of aggregated sentiment ratio (Positive/Negative) or (Negative/Positive) reaches a high level of more than 50%. Then they applied the FB-LDA model, which extracts all Foreground Topics, then it analyzes the documents which appeared earlier in the Background period. The model examines similarities between Foreground topics and Background topics, then finally extracts all Emerging Topic, which are the Foreground topics that did not show high similarity with Background topics.

Fig. 2 simplifies the Foreground-Background topic categorization task, where detected Emerging Topics are highlighted in green color. Final stage applies a Reason Candidate and Background LDA (RCB-LDA) model to extract the most representative tweet for each Emerging Topic.

Tan *et al.* [6] applied FB-LDA on the above mentioned SIGIR dataset which contains 924 abstracts from Information Retrieval subjects. The model managed to successfully handle the task of detecting Emerging Topics that appeared during the last three years, however, as indicated in Table 2, many old background topics were also presented by the model along with Emerging Topics.

In addition to the above-mentioned limitation, FB-LDA does not have clear guidelines for selecting the number of topics.

3) EVENT DETECTION

Event Detection method for analyzing sentiment variations was tailored by Jiang *et al.* [11] who were inspired by the Topic-Sentiment Mixture (TSM) models [12] to trace abrupt increases in document number for discussed topics, then correlate them with sentiment variations. Their framework

TABLE 2. FB-LDA results for SIGIR dataset [6].

	Research Topics	Top Words	
1	Exploiting Users' Behavior Data	Behavior Search Model User Click Log Session Data	✗
2	Probabilistic Factor Models for Recommendation	User Recommendation Person Interest Facet Factor Latent	✗
3	Search Result Diversification	Result Search Vertical Diverse Diversify Subtopic Show	✓
4	Query Suggestions	Query Search Suggest Engine Log Reformulation Predictor	✗
5	Quality of User-generated Content	Label Quality Book Crowdsource Select Flaw Impact Sample	✗
6	Twitter Stream Mining	Stream Twitter Context Tweet Entity Toponym Context-aware	✓
7	Image Search and Annotation	Image Visual Attribute Estimate Face Privacy Flickr Facial	✗
8	Search Result Cache Invalidation	Time Result Temporal Cache Evaluate Update Invalidate	✓
9	Temporal Indexing	Collect Index Time Web Structure Temporal Archive Time-travel	✓
10	Hashing for Scalable Image Retrieval	Retrieval Hash Example Code Method Propose Application	✓

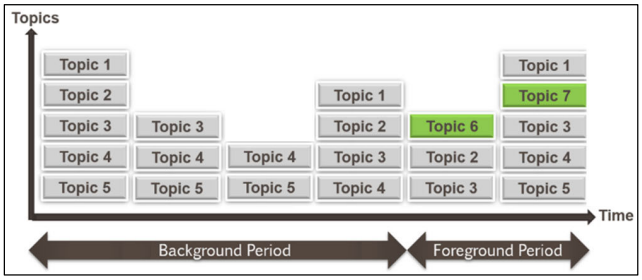

FIGURE 2. Emerging topics appear only in foreground documents.

is called Topic Sentiment Change Analysis (TSCA). It uses a rule-based method to extract sentiment, and Probabilistic Latent Semantic Analysis (PLSA) [13] to extract topics from text. Though the Event Detection method showed reasonably good results when topics are heavily discussed inside documents, it does not detect lower frequency topics which could be the main reasons for sentiment variations. [8].

B. SENTIMENT ANALYSIS FOR SHORT TEXTS

During the last four decades, many techniques were introduced to carry out the Sentiment Analysis task, which aims to detect subjectivity and polarity of texts at sentence-level, document-level, and aspect-level [14]. The 1980's witnessed significant research work on Sentiment Analysis, like analyzing subjective and objective texts [15], cognitive feature of sentiments [16], building affective lexicons [17]. In the 1990's, WordNet [18], Part of Speech (POS) Tagging [19], Parsing Trees based on Statistical Methods [20], [21], directional interpretation (positive/negative/neutral) of a given text [22], predicting the semantic orientation of adjectives [23], and Fuzzy Model [24] were introduced for data mining and used for sentiment analysis.

With the beginning of the 21st century, research work on Sentiment Analysis witnessed major enhancements. SentiWordNet [25] was published to provide a lexical resource like WordNet but dedicated for Sentiment Analysis, and Sentic Computing [26] was used to raise Sentiment Analysis to a new level. Machine Learning techniques became dominant in the Sentiment Analysis field. Majority of studies were carried out using Supervised Learning techniques [27], [28], However some Unsupervised Learning techniques [29] could also achieve good results. Bootstrapping method was presented to build lexicon of sentiments/subjectivity for languages that do not have enough resources [30].

Semi-Supervised Learning techniques [31] and Hybrid methods [32] were also used by some researchers and achieved good results.

Deep Learning techniques, including Recurrent Neural Networks (RNN), Convolutional Neural Networks (CNN), and Deep Neural Networks (DNN) have also showed excellent results when compared to other Machine Learning methods for Sentiment Analysis [33], especially when Word Embedding [34] representation is used with the Deep Learning algorithms. The Bidirectional Encoder Representations from Transformers (BERT) algorithm [35], which is a Neural Network-based technique, has also shown good results when applied on Sentiment Analysis. Fig. 3 summarizes the main techniques that have been used to handle Sentiment Analysis so far.

With the emergence of social media, additional challenges faced the Sentiment Analysis task as handling short texts requires special considerations. For instance, extracting sentiment from tweets through Supervised Learning methods would need significantly large annotated multi-domain datasets. As a result, Lexicon-based methods were found more efficient, so far, for handling short texts' Sentiment Analysis [36]. Three Lexicon-based tools are still being used frequently by various researchers to extract sentiment from short texts. These are SentiStrength [37], TextBlob [38], and VADER [39].

II. FILTERED-LDA FOR SENTIMENT REASONING

Inspired by the FB-LDA Model, we introduce a Filtered-LDA framework, which aims to overcome the four main limitations of FB-LDA by (1) Enhancing the topic categorization accuracy through ensuring low Perplexity Score for the model and applying multiple settings of LDA hyperparameters to perform a deep scan for discussed topics, (2) removing all documents that include old/background topics to ensure that final output will include Emerging Topics only, (3) enhance the interpretability of detected Emerging Topics by using the highest LDA Coherence Score and reducing the chance of using words from old/background topics, and (4) use accurate sentiment variation criteria.

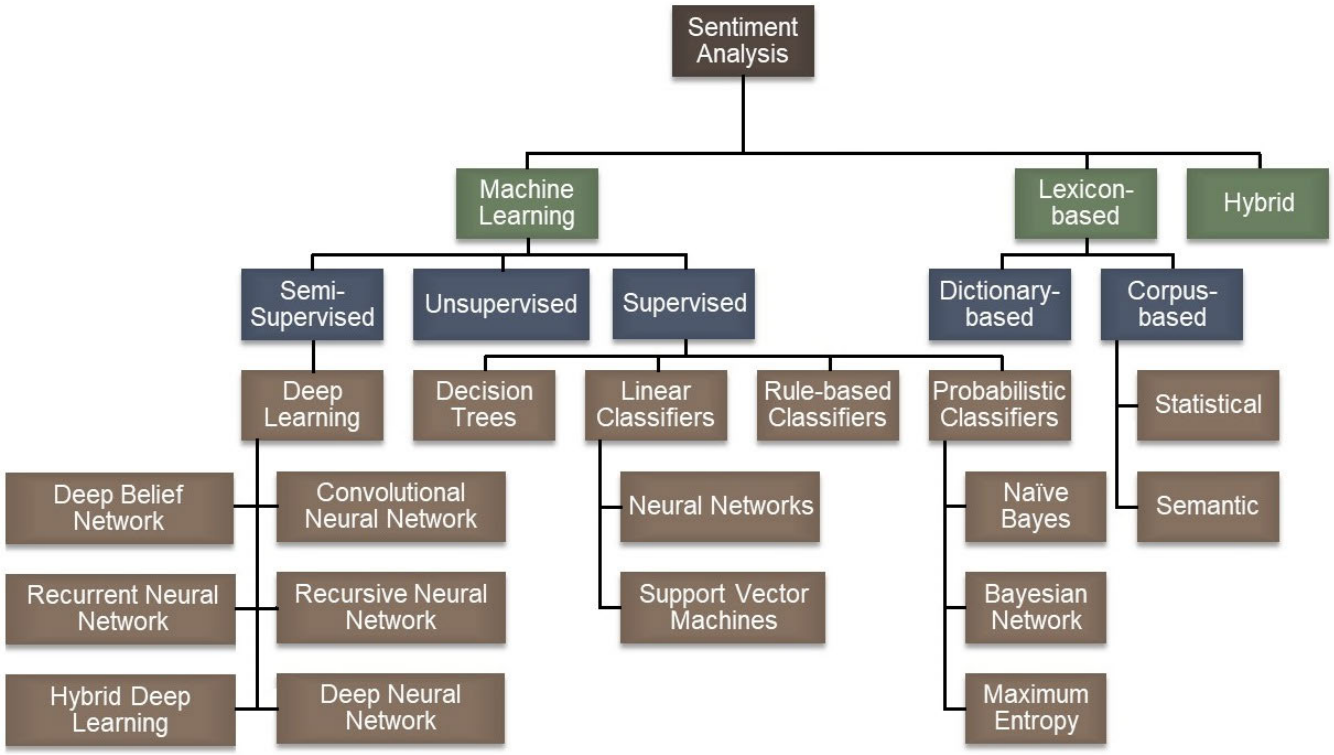

FIGURE 3. Main sentiment analysis techniques.

Given that our study focuses only on the Sentiment Reasoning task, we shall not propose a new method for extracting sentiment level. However, our experiment will compare available Sentiment Analysis tools to select highest performing classifiers. Our proposed Filtered-LDA framework can work with any Sentiment Analysis tool that produces acceptable sentiment classification results.

Fig. 4 shows the Filtered-LDA framework, which starts with a preprocessing step to normalize all tweets. Removal of stop words shall exclude negation words like "Not" to ensure correct sentiment classification for negated sentences.

Once Sentiment Analysis task is carried out, the system detects major sentiment variations. Tan *et al.* [6] used (POS/NEG) and (NEG/POS) ratios to monitor these variations, which may result in false detection of major variations when numbers of both positive and negative tweets are low, and when both positive and negative variation events occur during same period. Therefore, we shall use the variation measurement method which is proposed in [8], where the (POS/NEG) and (NEG/POS) peaks are combined with major increase in positive and negative sentiment levels, respectively. Once sentiment variation period is detected, all tweets during that period are labeled as Foreground tweets. The Background tweets are those appeared before the start of the Foreground period. Like Tan *et al.* [6], we shall extend the duration of the Background period to be double of the Foreground period to ensure that detected Emerging Topics are genuinely new. Longer Background periods can also be used. Low-frequency words shall be removed from Background tweets to reduce the chance of merging them with Emerging Topics.

The cleaned dataset is now ready for the first Topic Modeling process. To automatically select the number of topics "K1" that guarantees best Perplexity Score, a Hierarchical Dirichlet Processes (HDP) [40] is applied on the full cleaned dataset. Since the accuracy of HDP relies on its Term Weighting Scheme (TM) [41], the system compares the Coherence Scores of HDP using multiple TM, then it selects the TM that produces highest Coherence Score. The framework compares three TM outputs; these are the Inverse Document Frequency term weighting TM IDF, the Pointwise Mutual Information term weighting TM PMI, and the TM ONE which considers every term equal.

Detected HDP topics shall be sent to the framework output stage to complement the main system's output of Emerging Topics. HDP topics give the user an overall picture about discussed topics during Background and Foreground periods. All topics are ranked based on the number of tweets in which they appear as Dominant Topics. The output shall also show for each topic the most representative tweet wherein that topic has the highest probability. It shall also draw the curve of Topic Over Time to track the evolution of topic inside both Background and Foreground periods. However, this output shall be used for user support only as the clear demarcation of Emerging Topics is done after the Cascaded LDA block.

The concluded number of HDP topics "K1" is used for the Cascaded LDA block which can include high number of

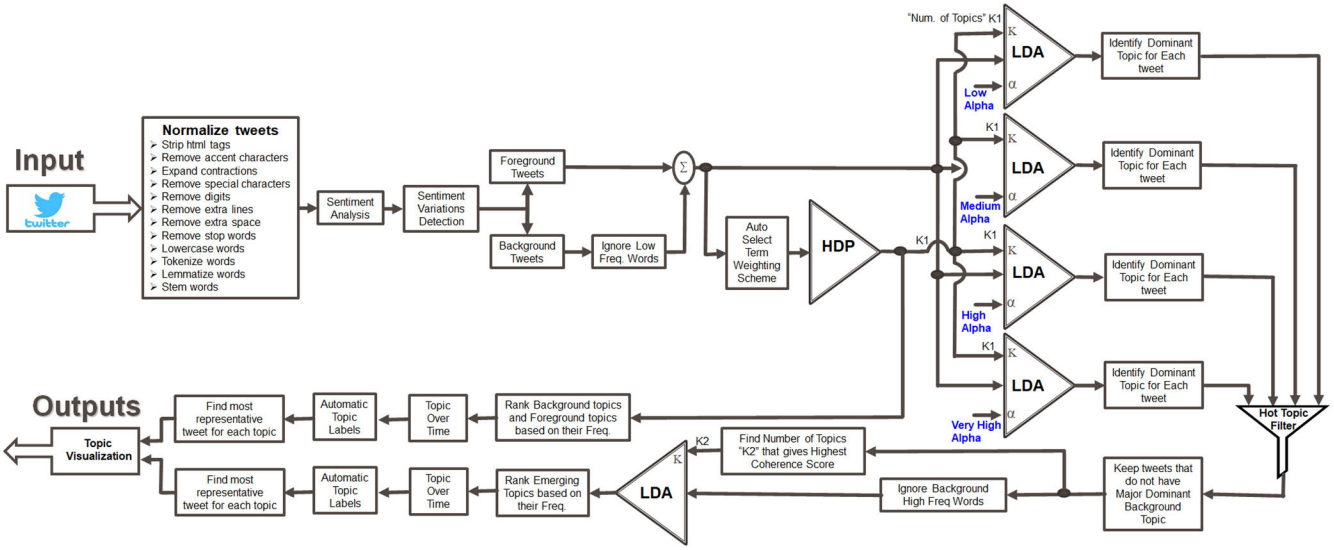

FIGURE 4. Sentiment reason mining framework for twitter.

LDA Models with different Alpha hyperparameters settings. In our framework, we used 4 LDA Models only as we could achieve good results with this number. Increasing the number of models further in the Cascaded-LDA block slows down the program speed, though it also enhances the Emerging Topic Detection performance. Alpha (α) hyperparameter of LDA determines combination of topics inside a tweet, whereas Beta (β) hyperparameter determines combination of words for each topic. For example, if you increase the value of Alpha, the combination of topics will increase [42]. Therefore, applying multiple Alpha hyperparameters ensures better scanning of the tweets as it emulates reading these tweets from multiple distances, which ensures better clustering for the topics.

Each model in the Cascaded LDA block is followed by a process of labeling each tweet by its topic that has the highest probability inside that tweet. That Dominant Topic will be used in the next step to decide whether that tweet belongs to Emerging Topics or Background Topics.

If a Dominant Topic's tweet appears more than once in the Cascaded LDA outputs, it shall be identified by the followed filter as an Emerging Topic's tweet. This ensures high confidence of the topic clustering process as the filtered tweet is categorized as an Emerging/Hot Topic's tweet by multiple values of Alpha hyperparameter. The Hot Topic filter defines the threshold of maximum number of Emerging Topic's tweets that can appear during the background period. This threshold is a variable setpoint, which is defined by the User. For instance, if this threshold is set to 0%, all Emerging Topics shall be those that did not appear in any Background tweet. In our experiments, we set this threshold to 5%. For instance, when a topic appears in 100 tweets during Foreground period, it shall be considered an Emerging Topic if it did not appear in more than 5 tweets during the Background period.

The output of the filter will be all Foreground tweets that are labeled by the system as Emerging Topic's tweets. These shall be applied to a final LDA Model that has high Coherence Score to ensure interpretability of LDA outputs by a human. The system uses multiple number of topics to automatically identify "K2" which produces the highest Coherence Score for the Topic Model. Final stage includes multiple forms of Topic Visualization to ensure easy human understanding of the Candidate Reasons for Sentiment Variations. Curves of topic frequency over time are drawn by counting the number of tweets wherein a topic is identified as Dominant Topic. Automatic topic labeling [43] is also used to select few keywords that represent each topic. Finally, the most representative tweet for each Emerging Topic is identified by selecting the tweet in which an Emerging Topic has the highest probability.

III. TWITTER DATASETS

To compare accuracies of Sentiment Analysis classifiers on tweets, we carried out experiments on two different datasets. First one is the US Airlines Twitter Dataset [44], which includes 14,640 tweets of customer feedbacks on six airlines. Each tweet is manually labeled for positive, neutral, or negative sentiment. This dataset serves as a Reference Dataset or a Gold Standard Dataset for comparing main classifiers.

We selected the US Airlines Twitter Dataset to train our classifiers because it includes Neutral sentiment label in addition to both Positive and Negative labels. Unfortunately, larger labeled Twitter datasets, like Sentiment140 Dataset [45] that includes 800,000 Positive and 800,000 Negative tweets, do not include Neutral tweets, which are important for our experiment as we shall exclude Neutral tweets from the Reason Mining task. Furthermore, we are not satisfied with the quality of Sentiment140 annotation as we could easily identify many annotation errors.

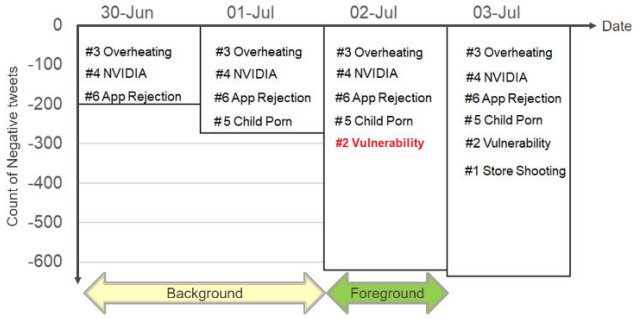

FIGURE 5. Ground truth dataset main topics.

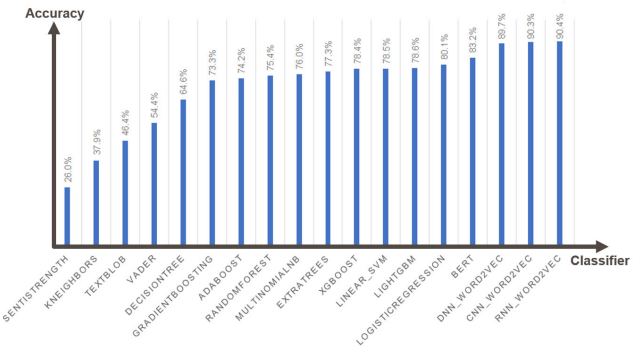

FIGURE 6. Accuracy of sentiment classifiers trained and tested on same twitter domain.

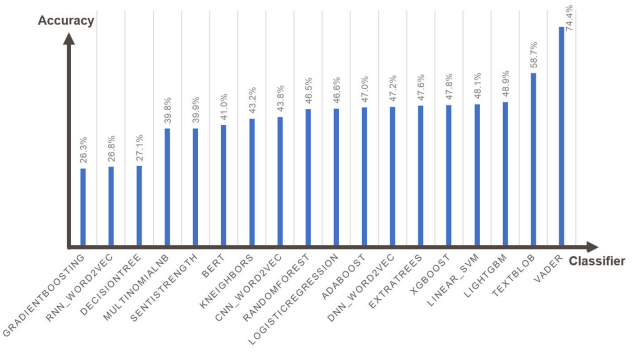

FIGURE 7. Accuracy of sentiment classifiers trained on one domain and tested on a different twitter domain.

To illustrate, these short tweets are annotated as Negative: "Yup", "*Me too", "I see", " At work", " almost bedtime", " currently at work", "I want the new GG episode already*", and "*I love you, Buck*".

Second dataset is the Stanford Twitter Dataset (STD-2009), which contains 476 million tweets from 1st of June 2009 to 31st of December 2009. It is estimated that STD-2009 includes around 20-30% of public tweets during the mentioned 7 months period [46].

To compare our Filtered-LDA results with the FB-LDA [6], we extracted all 643,264 English STD-2009 tweets that discuss "Apple", and 1,354,394 tweets that discuss "Obama".

For additional testing of Sentiment Analysis classifiers, we extracted a Ground Truth dataset from STD-2009 by manually labeling positive/neutral/negative sentiment and the reason of positive and negative sentiments for all the 5,082 tweets related to "Apple" from 30th June 2009 to 3rd of July 2009. The dataset includes 24.5% Negative, 40.6% Neutral, and 34.9% Positive tweets. It is used to compare accuracies of main Sentiment Analysis classifiers on our real-life dataset. The same dataset was used in [8] to demonstrate main shortcomings of existing Sentiment Reasoning methods. We shall use the annotated sentiment reasons to test the performance of our Filtered-LDA framework. Fig. 5 shows major variation on Negative sentiment level on 02nd of July 2009 when compared to the previous two days. It also shows the highest frequency topics which were discussed on each day of the Ground Truth Dataset, wherein the SMS Vulnerability topic is the Emerging Topic, which caused major sentiment variation.

IV. COMPARING SENTIMENT ANALYSIS CLASSIFIERS

To select the highest performing Sentiment Analysis classifier for our Twitter dataset, (1) we compare the accuracy of main classifiers on the US Airlines Twitter Dataset, and (2) we examine the consistency of these classifiers when the domains of training dataset and testing dataset are different by testing them on the Ground Truth dataset.

For text preprocessing and Sentiment Analysis classification, we used Python version 3.9 along with multiple packages, wrappers, and libraries, including NLTK, Spacy, Pandas, Numpy, Sklearn, Genism, Matplotlib, Torch, Transformers, Keras, Tensorflow, Sentistrength, TextBlob and VaderSentiment. Word2vec representation is used for Deep Learning algorithms to enhance classification accuracy [33]. It learns word embeddings by using a 2-layer Neural Network [34]. The text preprocessing stage excludes all negation terms from the stop-words-removal as these terms are important to conclude sentiment polarity. Emoticons are kept when VADER is applied because it is capable of assigning sentiment levels to both Emoticons and words. For other sentiment classifiers, Emoticons are automatically replaced by their Wikipedia meanings [47]. Fig. 6 shows the obtained accuracy for each classifier on US Airlines Dataset. For Learning-based algorithms, 90% of the tweets were used for training, and 10% for testing. Same figure shows results published in [33] for applying Deep Learning algorithms using Word2vec representation on US Airlines Dataset.

Unfortunately, all Learning-based algorithms show totally different results when they were re-tested on the Ground Truth Twitter dataset as shown in Fig. 7, where VADER provided highest accuracy, followed by TextBlob. Due to lack of large Twitter dataset with annotated positive, negative, and neutral sentiments so far, it is not possible to achieve high classification accuracy when Learning-based algorithms are trained on a small single-domain Twitter dataset and tested on a different domain.

In our experiments, many of these Learning-based algorithms produced excellent results when trained and tested

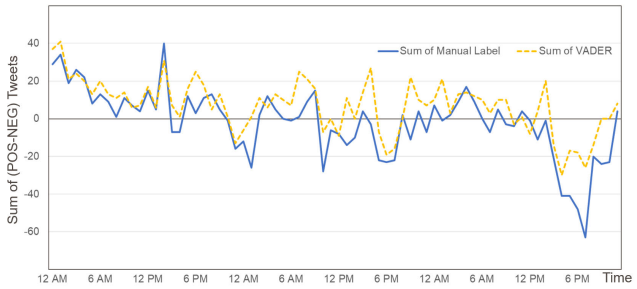

FIGURE 8. Hourly aggregated overall sentiment for ground truth dataset using VADER vs manual annotation.

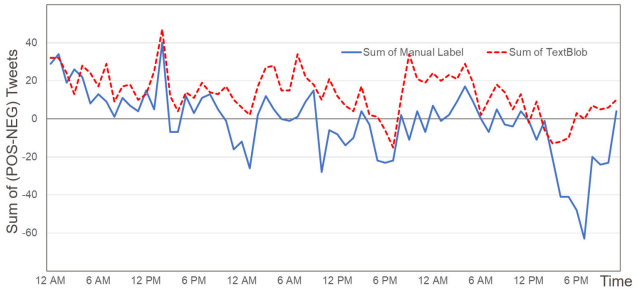

FIGURE 9. Hourly aggregated overall sentiment for ground truth dataset using TextBlob vs manual annotation.

on the same domain of US Airlines' customer feedbacks. However, when the same trained models were tested on the domain of Apple products, they failed to achieve high accuracies. Both VADER and TextBlob Lexicon-based methods produce more reliable outputs for Stanford Twitter Dataset (STD-2009).

Botchway *et al.* [48] compared accuracies of multiple Lexicon-based sentiment classifiers including VADER, TextBlob, SentiWordNet, and AFINN on Twitter, and they also concluded that VADER outperformed other Lexicon-based tools.

As explained in [39], VADER (Valence Aware Dictionary for sEntiment Reasoning) was developed to address sentiment classification challenges for social media texts. It employs a mixture of Rule-based and Lexicon-based approaches. VADER identifies common expressions, jargon, contractions, and terms. Furthermore, it accounts for grammatical structures, like negation, punctuation, prevarication, and exaggeration, which are commonly used on Twitter.

Due to its simple mechanism, VADER does not require a lot of computational resources, thus its speed is suitable for online Twitter processing. Moreover, unlike Learning-based algorithms, it does not need training, therefore consistency of its performance is not seriously impacted by the differences between domains of training and testing datasets. Hence, VADER shall be selected for our Filtered-LDA Sentiment Reasoning experiments.

Fig. 8 and Fig. 9 show the hourly aggregated overall sentiment for VADER and TextBlob respectively, when compared to the manually annotated sentiment for the Ground Truth dataset. VADER could emulate major positive and negative actual sentiment variations, whereas TextBlob looks biased to positive tweets.

V. FINDING REASON CANDIDATES

As VADER has been selected for carrying out the Sentiment Analysis part, the Filtered-LDA is now ready for analyzing STD-2009 tweets to interpret public sentiment variations related to the 643,264 tweets about "Apple" and 1,354,394 tweets about "Obama" from 1st of June 2009 to 31st of December 2009.

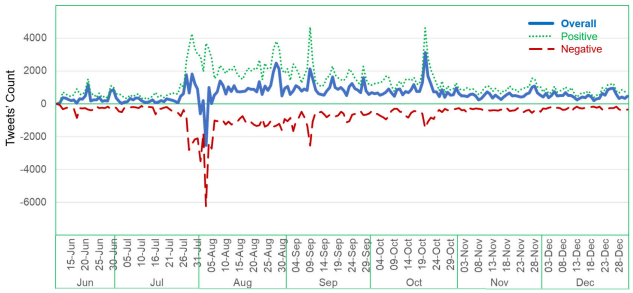

FIGURE 10. Daily aggregated "Apple" sentiments using VADER.

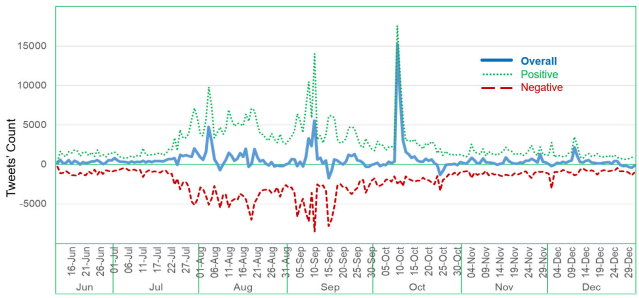

FIGURE 11. Daily aggregated "Obama" sentiments using VADER.

A. SENTIMENT VARIATION PERIODS

The system aggregates sentiment levels on daily basis by separately accumulating the number of positive tweets, negative tweets, and overall sentiment wight which is the sum of positive tweets' count minus negative tweets' count. If the user of the system is interested in monitoring sentiment variations for either shorter or longer periods, like hourly or weekly, then aggregation of tweets' counts shall be calculated accordingly. Fig. 10 and Fig. 11 show the sentiment curves of "Apple" and "Obama" respectively.

As shown earlier in Fig. 4, once the Sentiment Analysis is completed, positive and negative sentiment variation periods are identified. Table 3 shows all sentiment variation dates using the measurement criteria of Tan *et al.* [6] by identifying 50% peaks of (POS/NEG) and (NEG/POS) ratios. In the same table, we marked the correct sentiment variation dates using the criteria proposed in [8], where major increase in positive and negative sentiment levels are also considered in the measurement process..

TABLE 3. Sentiment variation dates using criteria applied in [6].

APPLE Pos/Neg VAR > 50%	APPLE Neg/Pos VAR > 50%	OBAMA Pos/Neg VAR > 50%	OBAMA Neg/Pos VAR > 50%
18-Jun-09	17-Jun-09 ✓	09-Oct-09 ✓	18-Aug-09
21-Jun-09 ✓	22-Jun-09	26-Oct-09	11-Oct-09
24-Jun-09	26-Jun-09	16-Nov-09	24-Oct-09 ✓
29-Jun-09 ✓	01-Jul-09	28-Nov-09 ✓	29-Nov-09
06-Jul-09 ✓	02-Jul-09 ✓	10-Dec-09 ✓	26-Dec-09 ✓
20-Jul-09	11-Jul-09 ✓		
26-Jul-09	16-Jul-09 ✓		
04-Aug-09	28-Jul-09 ✓		
06-Aug-09	01-Aug-09 ✓		
31-Aug-09	03-Aug-09 ✓		
10-Oct-09 ✓	30-Aug-09		
31-Oct-09	18-Sep-09		
05-Nov-09	22-Sep-09 ✓		
18-Nov-09	29-Sep-09		
	02-Nov-09 ✓		
	19-Dec-09		
	28-Dec-09 ✓		

Legend
✓ Correct Variation Dates
Correct Negative Variation
Correct Positive Variation

The unmarked dates in the table do not have major increase in positive or negative sentiment levels although the (POS/NEG) and (NEG/POS) ratios are high, which proves that our used criteria are more accurate. Only the marked dates are automatically detected by the framework, which identified these days as Foreground periods. The two days before each Foreground period are identified as Background periods.

B. EMERGING TOPIC DETECTION

Mallet [49] wrapper for Gensim [50] is used to apply LDA with optimized Gibbs Sampling [51] in our framework. We did not use standard Gensim LDA although it is faster because it employs Variational Bayes sampling method [52] which gave us lower Coherence Scores in our experiments. We used tomotopy toolkit [53] to apply HDP with Gibb Sampling and to utilize available Automatic Topic Label module [54].

For the Cascaded LDA, four different values of Alpha are used. Low ($\alpha = 1$), Medium ($\alpha = 50$), High ($\alpha = 100$), and Very High ($\alpha = 200$) values are selected to achieve zooming effect for LDA when analyzing tweets. Lower Alpha values ensure capturing topics when tweets are represented by a combination of few topics, whereas higher Alpha values capture topics when tweets are represented by a combination of more topics.

For the final LDA Model, we selected the highest frequency Emerging Topic to represent main Reason Candidate for each sentiment variation. Table 4 and Table 5 summarize the automatically detected Reason Candidate. The most representative tweet for each Reason Candidate is also identified through selecting the tweet wherein the Emerging Topic has highest probability. Such representation was proposed in [6], and it is useful for the user as it ensures better understanding of topics. This is convenient for the user because of the short length of tweets. For longer documents, we propose utilization of text summarization for the most representative documents, which can be simply implemented by available tools like Gensim Summarizer [55], which employs proposed TextRank algorithm's implementation method in [56], or the BERT Summarizer [57], which employs text summarization with Pretrained Encoders [58].

To verify concluded Reason Candidates, we manually examined the 78,672 tweets of all Foreground and Background periods. We fully agreed with all proposed reasons, though we did not agree with positive/negative sentiment classification for some tweets, which is expected because of the 74.5% accuracy of VADER for this dataset.

Unlike FB-LDA, the Filtered-LDA framework ensures detection of Emerging Topics only as it excludes all Background topics, and it applies better sentiment variation criteria to identify Foreground and Background periods. Hence, it is not a surprise that Filtered-LDA concluded more accurate reason candidates when compared to FB-LDA. Nevertheless, Tan *et al.* [6] introduced the Reason Candidate and Background LDA (RCB-LDA) Model to rank the candidate reasons. Table 6 shows an example of RCB-LDA results when applied to STD-2009 dataset.

Although RCB-LDA provides additional useful information for the user by showing the count of Reason Candidate tweets, it could not provide an accurate picture about the actual reason of negative sentiment variation towards Apple from 1st to 3rd of July 2009. As clear from Fig. 5, the actual spike of negative sentiment occurred on 2nd of July when the Emerging Topic of "SMS Vulnerability" appeared, whereas the Emerging Topic of "Store Shooting" started only on 3rd of July. Hence the actual reason of sentiment spike is the "SMS Vulnerability", which is detected successfully by the Filtered-LDA framework as shown in Table 4-b. Moreover, the mentioned RCB-LDA count of tweets for each reason candidate is also inaccurate. For instance, the actual count of "iPhone Overheating" is 27 tweets on 1st of July, 131 tweets on 2nd of July, and 92 tweets on 3rd of July, whereas RCB-LDA shows a total count of 179 tweets only. Furthermore, this topic of "iPhone Overheating" appeared earlier in 24 tweets on 30th of June, which means it is not an Emerging Topic at all. Filtered-LDA ranks the candidate reason topics based on the number of tweets in which an Emerging Topic is a Dominant Topic. It also ranks Background topics to provide a full picture for the user.

C. OVERALL SENTIMENT SPIKES

It is important to analyze positive and negative sentiment variations separately, however it is also useful to track the overall sentiment level as it forms a simple dashboard for public sentiment. If the aggregated sum of positive tweets is higher than negative tweets, the overall sentiment is positive, and vise versa. As shown earlier in Fig. 10 and Fig. 11, the overall sentiment levels of "Apple" and "Obama" show multiple positive and negative spikes throughout the shown period.

Fig. 12 marks all positive peaks of Apple's overall sentiment by a green circle whenever the level exceeds 3,000 tweets. It also marks all negative peaks of Apple's overall sentiment by red circles whenever the level exceeds -1,000 tweets. Fig.13 shows the same for Obama's overall sentiment. With this threshold, only 1 positive peak

TABLE 4. Filtered-LDA reason candidates for (a) positive and (b) negative "Apple" sentiment variations.

(a)

Positive Var > 50%	Main Reason	Most Representative Tweet
21-Jun	Unlock iPhone	iPhone Apple AT&T - How to jail break or unlock your iPhone , Break Free
29-Jun	Universal Phone Charger	Apple Agrees To Adopt Micro-USB Phone Charger In Europe
06-Jul	Micro Projectors	Cool!! : Apple to Add Micro Projectors to iPhone and iPod Touch
10-Oct	End of AT&T's iPhone Exclusivity	Why the End of AT&T's iPhone Exclusivity Would Be Good for Apple

(b)

Negative Var > 50%	Main Reason	Most Representative Tweet
17-Jun	Delayed launch of OS 3.0	BOOOOO! Apple has delayed the launch of iPhone OS 3.0 by a day: will now be released on Thursday
02-Jul	SMS vulnerability	Not good: Apple patching serious SMS vulnerability on iPhone
11-Jul	Google Stealing Apple's Ideas	Oooh - contentious! 'Why Google Is Stealing Apple's Ideas'
16-Jul	Stopping Hunters ads	Apple demanded Microsoft to stop its Laptop Hunters ads
28-Jul	Blocking Google Voice app	Apple Is Growing Rotten To The Core, Blocks Google Voice app from app store
01-Aug	FCC investigates app rejection	FCC seeks details on Google app rejection for iPhone. I side against Apple on this one without doubt.
03-Aug	Eric Schmidt Resigns	Eric Schmidt Resigns from Apple's Board of Directors
22-Sep	Stealing Microsoft Employees	RETAIL WAR: Microsoft Cherry Picking Apple Store Employees
02-Nov	Killing Hackintosh Netbook	Is Apple Trying to Kill the Hackintosh Netbook? Snow Leopard 10.6.2 Ditches Atom CPU Support
28-Dec	Dirty competition	Poor Pystar. Where's the healthy competition? If You Can't Beat Apple Sell T-Shirts [Revenge of Psystar!]

TABLE 5. Filtered-LDA reason candidates for (a) positive and (b) negative "Obama" sentiment variations.

(a)

Positive Var > 50%	Main Reason	Most Representative Tweet
09-Oct	Award of Nobel Peace Prize	President Obama has been awarded the 2009 Nobel Peace Prize for his "extraordinary" diplomatic efforts.
28-Nov	Vote for Presidency's 2nd term	do you want President Obama in for 2 terms? just vote - let the public know and get a $100 Visa giftcard
10-Dec	Obama's speech at Nobel.	President Obama is giving his #nobel peace prize speech right now I'm up next for my award in Chemistry

(b)

Negative Var > 50%	Main Reason	Most Representative Tweet
24-Oct	Swine flu	Obama declares swine flu a national emergency
26-Dec	Terror Attack	Obama Orders Heightened Security After Suspected Terror Attempt

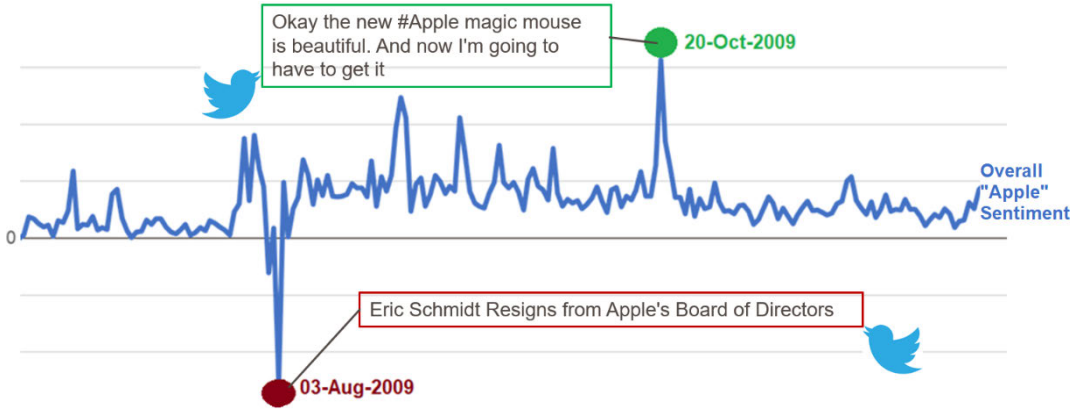

FIGURE 12. Overall sentiment spikes' reason candidates visualization for "Apple".

and 1 negative peak are detected for Apple, whereas 3 positive peaks and 2 negative peaks are detected for Obama.

For Apple, the overall negative peak happened on 3rd of August, when a NEG/POS variation was also there as shown in Table 4-b. Therefore the most representative tweet of the topic about ''Eric Schmidt Resignation'' is shown in Fig.12 linked to the overall negative peak.

The overall Apple's positive sentiment peak happened on 20th of October when there was no major POS/NEG variation as both positive and negative tweets experienced around 250% rise in their counts on that date. To understand the reasons of the positive rise on that date, Filtered-LDA is applied for the Foreground period of 20th of October, and Background period from 18th to 19th of October. The concluded main reason candidate is Apple's announcement about ''Magic Mouse''. As a result, the most representative tweet for the ''Magic Mouse'' topic is shown in Fig. 12 linked to the overall positive peak of 20th of October.

Similarly, for Obama tweets, Filtered-LDA is applied on the positive tweets to understand the main reasons of the 3 overall positive peaks, and applied on negative tweets to understand the mean reasons of the 2 overall negative peaks. Fig. 13 shows the most representative tweet for each main reason candidate linked to its associated peak. Obama's birthday on 4th of August was the main reason candidate for the first positive peak. The second positive peak happened when the Republican politician, Joe Wilson, apologized about his behavior during Obama's speech. Third positive peak happened when Obama was awarded Nobel Peace Prize.

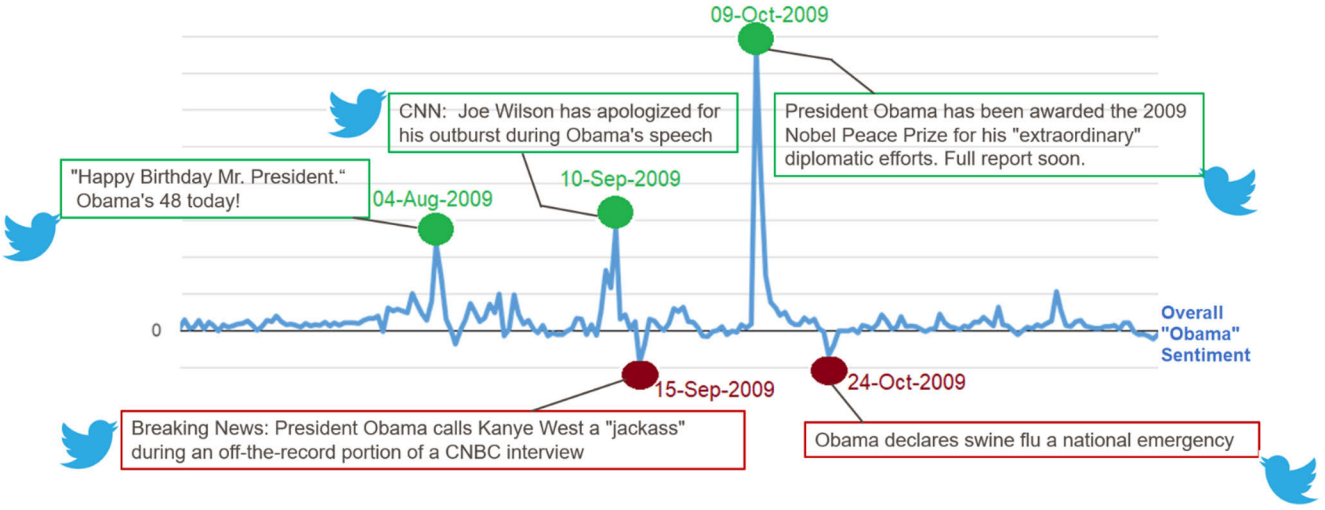

FIGURE 13. Overall sentiment spikes' reason candidates visualization for "Obama".

TABLE 6. RCB-LDA reason candidates for sentiment variation towards "Apple" from 1st to 3rd of July [6].

Cnt	Reasons
275	BREAKING Shooting at Arlington Apple Store! News Video via mashable. WTF.
191	Apple Patching Serious SMS Vulnerability on IPhone. Apple is Working to Fix an iPhone.
179	Apple warns on iPhone 3GS overheating risk.
101	Apple may drop NVIDIA chips in Macs following contract fight.
87	Child Porn Is Apple's Latest iPhone Headache.
84	App Store Rejections: Apple rejects iKaraoke app then files patent for karaoke player.

First negative peak happened when Obama mentioned the word ''jackass'' about the American rapper, Kanye West, during an interview. Second negative peak was caused by Obama's announcing Swine Flu a national emergency.

D. RELATIONSHIPS BETWEEN TOPICS

Sometimes, the sentiment variation reason candidates are casually linked. For example, on 1st of August 2009, the Federal Communications Commission (FCC) investigated Apple's rejection of Google Voice App. This event caused a negative sentiment variation for Apple on that date as shown in Table 4-b. After two days, on 3rd of August 2009, the CEO of Google, Dr. Eric Schmidt, resigned from Apple's board of directors. This event caused another negative sentiment variation for Apple on that date as shown in Table 4-b.

Some special Topic Models can be used to link topic together so that the user may have better understanding of the reason candidates and possible relationship between them. We used both Hierarchical Pachinko Allocation (HPA) [59] and Multi Grain Latent Dirichlet Allocation (MG-LDA) [60] to investigate the relationships between Reason Candidates. For instance, Fig. 14 shows outputs of both HPA and MG-LDA when they were separately applied on Apple tweets from 10th of July to 10th of August 2009. The output of HPA models suggests relationship between the Super-topic of ''Eric Schmidt Resignation'' and the Sub-topic of ''FCC & Google App Rejection''. A similar relation is also detected

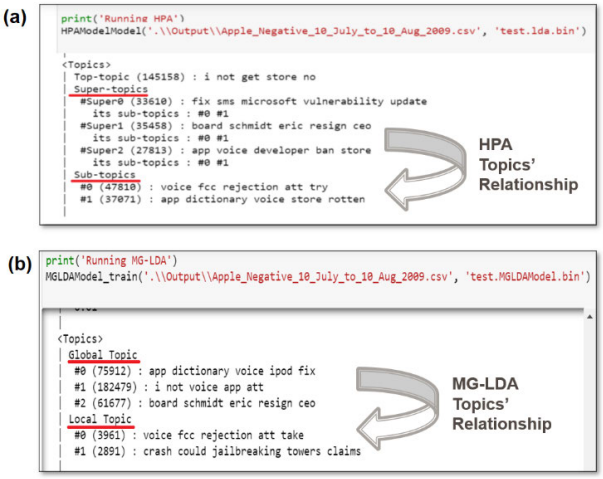

FIGURE 14. (a) HPA and (b) MG-LDA outputs for one month of Apple tweets from 10th July 2009.

by MG-LDA. We used tomotopy toolkit's HPA and MG-LDA functions to apply both models.

VI. CONCLUSION

This paper addressed the problem of automatically detecting reasons behind major sentiment variations on Twitter. It reviewed existing methods and identified their major shortcomings. To overcome these, we proposed a novel Filtered-LDA framework that could outperform base methods. It uses a Cascaded LDA block with multiple LDA hyperparameter values to zoom inside and outside texts for detecting Emerging Topics.

Our framework applies an enhanced sentiment variation measurement to detect changes on sentiment levels accurately. The outputs of the framework include conventional HDP topics and the Filtered-LDA Emerging Topics separately.

Visualization of topics include Topic Over Time curves, automatic topic labels, and most representative document for each topic. HPA and MG-LDA are then used to investigate possible relations between Reason Candidates.

The peaks of overall sentiment are identified by the system automatically. Finally, we proposed a novel dashboard, which links the most represented tweet of main reason candidates with their associated positive or negative sentiment spike.

Our experiments include comparison of various Sentiment Analysis classifiers on short texts. Although some Learning-based classifiers showed high accuracy when they were trained and tested on the same domain, they could not achieve good results when they were tested on a different domain. As a result, we selected VADER tool for our framework because it showed highest Sentiment Analysis accuracy for our dataset.

In our future work, we shall apply Filtered-LDA framework on Arabic tweets to check its performance for other languages. We shall also investigate other methods of monitoring sentiment by considering the importance of expressed opinion inside each tweet, which depends on number of retweets and number of followers of the tweet's author.

REFERENCES

[1] P. Tighe, R. Goldsmith, M. Gravenstein, H. Bernard, and R. Fillingim, "The painful tweet: Text, sentiment, and community structure analyses of tweets pertaining to pain," *J. Med. Internet Res.*, vol. 17, no. 4, pp. 1–19, 2015.

[2] K. Shakhovska, N. Shakhovska, and P. Vesely, "The sentiment analysis model of services Providers' feedback," *Electron. J.*, vol. 9, no. 11, pp. 1–15, 2020.

[3] H. Yin, S. Yang, and J. Li, "Detecting topic and sentiment dynamics due to COVID-19 pandemic using social media," in *Proc. Int. Conf. Adv. Data Mining Appl. (ADMA)*, Foshan, China, 2020.

[4] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *Proc. 23rd Int. Conf. Mach. Learn. (ICML)*, 2006.

[5] 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.

[6] 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.

[7] A. Goldstone, S. Galán, C. L. Lovin, A. Mazzaschi, and L. Whitmore, "An interactive topic model of signs," *Signs J.*, 2014.

[8] F. Alattar and K. Shaalan, "A survey on opinion reason mining and interpreting sentiment variations," *IEEE Access*, vol. 9, pp. 9636–9655, 2021.

[9] A. Giachanou, I. Mele, and F. Crestani, "Explaining sentiment spikes in Twitter," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, Indianapolis, IN, USA, Oct. 2016.

[10] 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)*, Karlskrona, Sweden, Aug. 2014.

[11] Y. Jiang, W. Meng, and C. Yu, "Topic sentiment change analysis," in *Proc. Int. Workshop Mach. Learn. Data Mining Pattern Recognit. (MLDM)*, Berlin, Germany, 2011.

[12] Q. Mei, X. Ling, M. Wondra, H. Su, and C. X. Zhai, "Topic sentiment mixture: Modeling facets and opinions in Weblogs," in *Proc. 16th Int. Conf. World Wide Web*, Banff, AB, Canada, 2007.

[13] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, Berkeley, CA, USA, 1999.

[14] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, no. 1, pp. 1–135, 2008.

[15] A. Banfield, *Unspeakable Sentences: Narration and Representation in the Language of Fiction*. New York, NY, USA: Law Book Co of Australasia, 1982.

[16] T. Winograd, *Language As a Cognitive Process: Syntax*, vol. 1. Lansing, MI, USA: Addison Wesley, 1983.

[17] G. Clore, A. Ortony, and M. Foss, "The psychological foundations of the affective lexicon," *J. Personality Social Psychol.*, vol. 53, no. 4, pp. 751–766, 1987.

[18] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to WordNet: An on-line lexical database," *Int. J. Lexicogr.*, vol. 3, no. 4, pp. 235–244, 1990.

[19] E. Brill, "Some advances in transformation-based part of speech tagging," in *Proc. 12th Nat. Conf. Artif. Intell. (AAAI)*, Seattle, WA, USA, 1994.

[20] S. Abney, "Statistical methods and linguistics," in *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. Cambridge, MA, USA, 1996.

[21] C. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999.

[22] P. Jacobs, Ed., "Direction-based text interpretation as an information access refinement," in *Text-Based Intelligent Systems: Current Research and Practice in Information Extraction and Retrieval*. Berkeley, CA, USA: Lawrence Erlbaum Associates, 1992, pp. 257–274.

[23] V. Hatzivassiloglou and K. McKeown, "Predicting the semantic orientation of adjectives," in *Proc. 35th Annu. Meeting Assoc. Comput. Linguistics, 8th Conf. Eur. Chapter Assoc. Comput. Linguistics*, Stroudsburg, PA, USA, 1997.

[24] R. Kruse, D. Nauck, and C. Borgelt, "Data mining with fuzzy methods: Status and perspectives," in *Proc. 7th Eur. Congr. Intell. Techn. Soft Comput. (EUFIT)*, Aachen, Germany, 1999.

[25] A. Esuli and F. Sebastiani, "SentiWordNet: A publicly available lexical resource for opinion mining," in *Proc. 5th Conf. Lang. Resour. Eval. (LREC)*, Genova, Italy, 2006.

[26] E. Cambria and A. Hussain, *Sentic Computing Techniques, Tools and Applications*. New York, NY, USA: Springer, 2012.

[27] A. Kumar and T. M. Sebastian, "Sentiment analysis: A perspective on its past, present and future," *Int. J. Intell. Syst. Appl.*, vol. 4, no. 10, pp. 1–14, Sep. 2012.

[28] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Philadelphia, PA, USA, 2002.

[29] P. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Philadelphia, PA, USA, 2002.

[30] C. Banea, E. Mihalcea, and J. Wiebe, "A bootstrapping method for building subjectivity lexicons for languages with scarce resources," in *Proc. 6th Int. Conf. Lang. Resour. Eval. (LREC)*, Marrakech, Morocco, 2008.

[31] M. K. Dalal and M. A. Zaveri, "Semisupervised learning based opinion summarization and classification for online product reviews," *Appl. Comput. Intell. Soft Comput.*, vol. 2013, pp. 1–8, Jan. 2013.

[32] B. Lu and B. K. Tsou, "Combining a large sentiment lexicon and machine learning for subjectivity classification," in *Proc. Int. Conf. Mach. Learn. Cybern.*, Qingdao, China, Jul. 2010.

[33] N. C. Dang, M. N. Moreno-Garcia, and F. D. L. Prieta, "Sentiment analysis based on deep learning: A comparative study," *Electron. J.*, vol. 9, no. 483, pp. 1–29, 2020.

[34] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. Int. Conf. Learn. Represent.*, Scottsdale, AZ, USA, 2013.

[35] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, Minneapolis, MN, USA, 2019.

[36] A. Giachanou and F. Crestani, "Like it or not: A survey of Twitter sentiment analysis methods," *ACM Comput. Surveys*, vol. 49, no. 2, pp. 28–41, 2016.

[37] 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.

[38] S. Loria. (2020). *TextBlob: Simplified Text Processing*. Accessed: Mar. 26, 2021. [Online]. Available: https://textblob.readthedocs.io/

[39] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. 8th Int. Conf. Weblogs Social Media (ICWSM)*, Oxford, U.K., 2014.

[40] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Sharing clusters among related groups: Hierarchical Dirichlet processes," in *Proc. 17th Int. Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2004.
[41] A. Wilson and P. Chew, "Term weighting schemes for latent Dirichlet allocation," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, Los Angeles, CA, USA, 2010.
[42] J. Hansen, "Inside latent Dirichlet allocation: An empirical exploration," *Knowl. Inf. Syst.*, pp. 1–21, 2016.
[43] Q. Mei, X. Shen, and C. Zhai, "Automatic labeling of multinomial topic models," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, San Jose, CA, USA, 2007.
[44] Crowdower. (Feb. 2015). *Twitter US Airline Sentiment*. Accessed: Jan. 30, 2021. [Online]. Available:
[45] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," Stanford, CA, USA, Tech. Rep., 2009.
[46] 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 (WWW)*, Raleigh, NC, USA, 2010.
[47] Wikipedia. *List of Emoticons*. Accessed: Apr. 1, 2021. [Online]. Available: https://en.wikipedia.org/wiki/List_of_emoticons
[48] R. Botchway, A. Jibril, Z. Oplatkova, and M. Chovancova, "Deductions from a Sub-Saharan African Bank's Tweets: A sentiment analysis approach," *Cogent Econ. Finance*, vol. 8, no. 1, pp. 1–19, 2020.
[49] A. McCallum. (2002). MAchine Learning for LanguagE Toolkit Website. University of Massachusetts Amherst. Accessed: Apr. 2, 2021. [Online]. Available: http://mallet.cs.umass.edu/
[50] R. Rehurek. (Apr. 2021). Gensim 4.0.1 Python Library for Topic Modelling. Python Software Foundation. Accessed: Apr. 2, 2021. [Online]. Available: https://pypi.org/project/gensim/
[51] L. Yao, D. Mimno, and A. McCallum, "Efficient methods for topic model inference on streaming document collections," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Paris, France, Jun. 2009.
[52] M. Hoffman and D. Blei, "Online learning for latent Dirichlet allocation," in *Proc. Adv. Neural Inf. Process. Syst., 24th Annu. Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2010.
[53] Bab2min. (Mar. 27, 2021). Tomotopy 0.11.1 Python Extension. Python Software Foundation. Accessed: Apr. 2, 2021. [Online]. Available: https://pypi.org/project/tomotopy/
[54] Bab2min. (Mar. 27, 2021) Module Tomotopy.Label. Accessed: Apr. 2, 2021. [Online]. Available: https://bab2min.github.io/tomotopy/v0.11.1/en/label.html
[55] *Gensim Summarization*. Accessed: Apr. 2, 2021. [Online]. Available: https://radimrehurek.com/gensim_3.8.3/summarization/summariser.html
[56] F. Barrios, F. Lopez, L. Argerich, and R. Wachenchauzer, "Variations of the similarity function of TextRank for automated summarization," in *Proc. Argentine Symp. Artif. Intell. (ASAI)*, Rosario, Argentina, 2015.
[57] K-Tahiro. (Mar. 30, 2021). Bert-Summarizer 0.1.4. Python Software Foundation. Accessed: Apr. 2, 2021. [Online]. Available: https://pypi.org/project/bert-summarizer/
[58] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," in *Proc. Conf. Empirical Methods Natural Lang. Process., 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, Hong Kong, 2019.
[59] D. Mimno, W. Li, and A. McCallum, "Mixtures of hierarchical topics with Pachinko allocation," in *Proc. 24th Int. Conf. Mach. Learn.*, Corvalis, OR, USA, 2007.
[60] I. Titov and R. McDonald, "Modeling online reviews with multi-grain topic models," in *Proc. 17th Int. Conf. World Wide Web (WWW)*, Beijing, China, 2008.

FUAD ALATTAR was born in Amman, Jordan. He is currently working as the Senior Vice President with Siemens, where he is managing the Digital Enterprise Services Business Unit in the Middle East. He has more than 25 years' experience in automation engineering, digitalization, data analytics, project management, and industrial cybersecurity fields, including the positions of the general manager, the operations manager, the sales director, and the engineering manager with main automation contracting (MAC) companies. He was selected by ARAMCO Standards' Department to carry out value engineering for critical process automation and protection systems. He has been an active researcher in signal Processing, machine learning, text engineering, 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 also 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. He founded and led the NLP Research Group and successfully secured internal and external funds and conducted research with international collaborators. 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, educational technology, and computational linguistics in particular Arabic natural language processing (NLP). He acts as the chair of international conferences, a member of the editorial board of reputed journals, and a book editor, a keynote speaker, and an external member of promotions committees.