Artificial Intelligence Frameworks for Sentiment Variations' Reasoning and Emerging Topic Detection

أنظِمَةُ ذَكاءٍ اصْطِناعي لتَفسيرِ تَغيُّرِ وِجْهات النَظَرِ واكتِشافِ المَوضوعاتِ الجَديدَة

by

FUAD ABDELWAHAB ABDELQADER ALATTAR

A thesis submitted in fulfilment

of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE

at

The British University in Dubai

July 2021

Artificial Intelligence Frameworks for Sentiment Variations’ Reasoning and Emerging Topic Detection

أنظِمَةُ ذَكاءٍ اصْطِناعي لتَفسيرِ تَغيُّرِ وِجْهات النَظَرِ واكتِشافِ المَوضوعاتِ الجَديدَة

by

FUAD ABDELWAHAB ABDELQADER ALATTAR

A thesis submitted to the Faculty of Engineering & IT in fulfilment of the requirements for the degree of DOCTOR OF PHILOSOPHY (PhD)

at

The British University in Dubai

July 2021

Thesis Supervisor

Prof. Khaled Shaalan

Approved for award:

__________________________	__________________________
Name	Name
Designation	Designation
__________________________	__________________________
Name	Name
Designation	Designation

Date: 15th July 2021

DECLARATION

I warrant that the content of this research is the direct result of my own work and that any use made in it of published or unpublished copyright material falls within the limits permitted by international copyright conventions.

I understand that a copy of my research will be deposited in the University Library for permanent retention.

I hereby agree that the material mentioned above for which I am author and copyright holder may be copied and distributed by The British University in Dubai for the purposes of research, private study or education and that The British University in Dubai may recover from purchasers the costs incurred in such copying and distribution, where appropriate.

I understand that The British University in Dubai may make a digital copy available in the institutional repository.

I understand that I may apply to the University to retain the right to withhold or to restrict access to my thesis for a period which shall not normally exceed four calendar years from the congregation at which the degree is conferred, the length of the period to be specified in the application, together with the precise reasons for making that application.

____

Signature of the student

COPYRIGHT AND INFORMATION TO USERS

ABSTRACT

Utilizing Sentiment Analysis techniques to monitor public opinions on social media has been an essential yet challenging task in the field of Artificial Intelligence (AI). Many studies were conducted during the last two decades to help users tracking public sentiments about entities, products, events, or other targets. However, these techniques focus on extracting overall positive/negative/neutral polarity of texts without identifying the main reasons for extracted sentiments. This thesis contributes to the very few studies that took one step ahead by developing novel models to understand what causes sentiment changes over time. Obviously, identifying main reasons for public reactions is valuable to decision-makers so that they can take necessary actions in a timely manner.

To develop our approach, we first examined existing Sentiment Reason Mining methods to identify their limitations, then we introduced our Filtered Latent Dirichlet Allocation (Filtered-LDA) Model that overcomes major deficiencies of base methods. This model can be used for multiple applications, including detection of new research trends from large sets of scientific papers, discovery of hot topics on social media, comparison of customer reviews for two products to identify their strong/weak aspects, and our focus topic of interpreting public sentiment variations.

The Filtered-LDA Model utilizes a novel Emerging Topic Detection technique for which we developed multiple AI frameworks. It emulates human approach for discovering new topics from a large set of documents. A human would first skim through all old and new documents to isolate the new ones that may contain Emerging Topics. These clustered documents are then analyzed to identify the high-frequency emerging topics. With this simple method, the impact of clustering errors is significantly reduced as the wrongly clustered documents do not usually contain main keywords of high-frequency emerging topics. Furthermore, the new frameworks introduce measures to genuinely reduce chances of detecting old topics and visualize candidate reasons online.

Given that some social media platforms, like twitter, use short-text documents, we first compared accuracies of state-of-the-art Sentiment Analysis classifiers to select the best performer for short-texts. Subsequently, the selected classifier is applied on a real-life large Twitter dataset, which includes around two million tweets, to extract positive/negative/neutral sentiments. The Filtered-

LDA Model is tested first on a Ground Truth dataset to validate that it outperforms baseline models, then it is finally applied on the large Twitter dataset to automatically conclude main reasons for sentiment variations

ملخص

يعتبر استخدام تقنيات تحليل المشاعر لمراقبة الرأي العام على وسائل التواصل الاجتماعي وسيلة هامّة محفوفة بالتحديات في مجال الذكاء الاصطناعي. وقد تم إجراء العديد من الدراسات خلال العقدين الماضيين لمساعدة المهتمين بمراقبة الرأي العام حول الشخصيات أو المنتجات أو الأحداث أو غيرها من الموضوعات. لكن تلك التقنيات تهتم فقط بتصنيف النصوص حسب المشاعر الإيجابية أو السلبية أو الحيادية الموجودة فيها ولا تتجاوز ذلك لمعرفة الأسباب وراء تلك المشاعر. لهذا فإن هذه الأطروحة هي مساهمة في الدراسات القليلة جداً التي أخذت البحث العلمي خطوة إلى الأمام من خلال تطوير تقنيات جديدة لفهم أسباب تغيّر وجهات نظر ومشاعر الجمهور مع مرور الوقت. فتحديد الأسباب الرئيسية لردود الفعل العامة الإيجابية أو السلبية هو أمر مهمّ لصناع القرار حيث يساعد على اتخاذ الإجراءات اللازمة في الوقت المناسب.

قبل تطوير التقنية الجديدة لتحديد أسباب تغيّر وجهات النظر قمنا في هذه الأطروحة باختبار الطرق الموجودة حالياً وذلك لمعرفة مدى كفاءتها، بعدها قدّمنا نموذجاً تصميمياً لمعالجة النصوص يقوم بتصفية توزيع دِرِشْلِيهْ الاحتمالي الكامن وذلك للتغلب على أوجه القصور الرئيسية التي تعاني منها الطرق الحالية. يمكن استخدام هذا النموذج التصميمي المبتكر في العديد من التطبيقات مثل معرفة أحدث توجّهات البحث العلمي من خلال معالجة نصوص مجموعة كبيرة من الأوراق والنشرات العلمية، وأيضاً اكتشاف الموضوعات الجديدة الأكثر تداولاً على وسائل التواصل الاجتماعي، ومقارنة آراء الزبائن حول مُنتَجَين لتحديد جوانب القوة والضعف في أيّ منهما، وكذلك التطبيق المتعلّق بأطروحتنا وهو تفسير تغيّر وجهات النظر والمشاعر العامة.

مبدأ عمل النموذج التصميمي الجديد يرتكز على تقنية مبتكرة لاكتشاف الموضوعات الناشئة، وقد عَرَضت هذه الأطروحةُ عدّة طرق لتنفيذها باستخدام الذكاء الاصطناعي. النموذج يحاكي أسلوب البشر لاكتشاف موضوعات جديدة في مجموعة كبيرة من الوثائق أو المستندات. حيث يقوم الإنسان أولاً بالخطوة الأولى وهي تصفّح سريع لجميع النصوص القديمة والحديثة بهدف عزل النصوص الحديثة التي يظنّ أنها تحتوي على موضوعات ناشئة. بعدها يقوم بالخطوة الثانية وهي قراءة النصوص المعزولة بتمعّن لتحديد الموضوعات الناشئة الأكثر تكراراً. باستخدام هذه الطريقة البسيطة يتمّ تقليل أثر أخطاء خطوة عزل النصوص بشكل كبير لأن النصوص التي تمّ عزلها عن طريق الخطأ لا تحتوي عادةً على كلمات هامّة تتعلّق بموضوعات ناشئة متكررة. كما أنّ النموذج الجديد يأخذ في عين الاعتبار عدّة تدابير تصميمية لتقليل فُرَص استنتاج موضوعات قديمة بالخطأ، وكذلك يقدّم طريقة عرض معلومات لتسهيل فهم الأسباب التي يستنتجها نظام الذكاء الاصطناعي لتفسير تغيّر وجهات النظر.

نظرًا لأنّ منصات تواصل اجتماعي مثل تويتر تتضمن نصوصاً قصيرة فقد قمنا أولاً بمقارنة دقة أحدث أنظمة تحليل المشاعر لاختيار أكثرها كفاءة في معالجة النصوص القصيرة. بعد ذلك قمنا بتطبيق النظام المختار على مجموعة كبيرة من التغريدات الحقيقية على موقع تويتر تضمّ حوالي مليوني تغريدة لتحديد المشاعر الإيجابية والسلبية والحيادية الموجودة فيها. وقبل تطبيق نموذجنا التصميمي الجديد قمنا باختبار دقته على مجموعة أصغر من نفس تلك التغريدات للتحقق من أنه يتفوق على النماذج الأخرى المستخدمة حالياً. بعدها طبّقنا النظام المبتكر على مجموعة التغريدات الكبيرة كاملة لاستنتاج الأسباب الرئيسية لتغيّر وجهات النظر والمشاعر تلقائيًا.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	**Meaning**
ABSA	Aspect-Based Sentiment Analysis
ASEM	Aspects and Sentiment of Events from Microblog
BERT	Bidirectional Encoder Representations from Transformer
BoW	Bag of Words
CNN	Convolutional Neural Networks
CRF	Conditional Random Field
DNN	Deep Neural Networks
DTM	Dynamic Topic Model
FB-LDA	Foreground and Background Latent Dirichlet Allocation
Filtered-LDA	Filtered Latent Dirichlet Allocation
HDP	Hierarchical Dirichlet Process
HMM	Hidden Markov Model
HPA	Hierarchical Pachinko Allocation
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSI	Latent Semantic Indexing
MALLET	MAchine Learning for LanguagE Toolkit
MG-LDA	Multi-Grain Latent Dirichlet Allocation
NB	Naive Bayes
NEG	Number of Negative Sentiment Tweets
NMF	Non-Negative Matrix Factorization
PLSA	Probabilistic Latent Semantic Analysis
PLSI	Probabilistic Latent Semantic Indexing
POS	Number of Positive Sentiment Tweets
RCB-LDA	Reason Candidate and Background LDA
RNN	Recurrent Neural Networks
RQ	Research Question
SIGIR	Special Interest Group on Information Retrieval

STD	Stanford Twitter Dataset
SVM	Support Vector Machines
TF-IDF	Term Frequency–Inverse Document Frequency
TM	Term Weighting Scheme
TPP	Time Period Partition
TSCA	Topic Sentiment Change Analysis
TSM	Topic-Sentiment Mixture
VADER	Valence Aware Dictionary for sEntiment Reasoning

Chapter 1 : Introduction

1.1 Research Problem

Social media platforms are certainly rich sources of public opinions on various types of topics. For instance, successful companies persistently adapt their management decisions and choices to increase the satisfaction index of their customers, therefor tracking response of their product consumers on social media is an important tool for sensing market's reaction around-the-clock. Thus, many software applications have been made available in the market to monitor sentiments of social media users. Such applications utilize Sentiment Analysis, which uses Artificial Intelligence techniques to extract polarity and subjectivity of texts.

Through Sentiment Analysis, it became feasible to watch online variations of positive, negative, and neutral sentiment levels for various types of targets, like celebrities, politicians, brand names, etc. However, there is an obvious additional task beyond Sentiment Analysis, which is learning the main reasons behind public sentiment variations. Currently, users of Sentiment Analysis applications need to manually analyze available trends and dashboards to understand what causes changes to sentiment levels.

As a result, there is a need for Sentiment Variations' Reason Mining models to automatically conclude possible reason candidates, which could be certain product features, manufacturing defects, rumors, national disasters etc. However, so far, such task has been addressed by only few studies, which suffer critical limitations. Moreover, due to such limited number of research projects, there is a strong need for assessing existing hypothesis on identifying explicit reasons for social media's sentiment variations. This makes the problem of taking Sentiment Analysis to the next level - by developing techniques that can efficiently detect sentiment variations and extract their reasons – a real challenge. The need for tackling this tricky problem was a strong motivation for writing this thesis.

1.2 Research Questions*

The next chapters of this thesis address the main research questions related to Sentiment Variations' Reason Mining. To do that, some real-life datasets shall be examined to understand behaviors of social media users when they express their opinions on various types of targets. Furthermore, experiments shall be conducted to identify existing research gaps and building the foundation of a new model that can overcome existing limitations. These are the main research questions:

* RQ (1) to RQ (9) are derived from the article "Alattar, F. & Shaalan, K., 2021. A Survey on Opinion Reason Mining and Interpreting Sentiment Variations. IEEE Access, Volume 9, pp. 39636-39655".

- **RQ (1) Explicit Reasons**: Do most of subjective tweets explicitly indicate reasons for 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 for the spike always have the highest topic frequency?
- **RQ (4) Events**: Can the Event Detection method always discover reasons for public sentiment variations?
- **RQ (5) Emerging Topic**: Is Emerging Topic Detection efficient for interpreting 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's reason always cause a spike in the overall sentiment level?
- **RQ (8) Topic Modeling**: Can conventional Topic Modeling methods help us understand reasons for sentiment variations?
- **RQ (9) Foreground-Background Topics**: Can the FB-LDA Model discover Emerging Topics within a sentiment variation period?
- **RQ (10) Efficient Topic Detection Technique**: How can Emerging Topics be efficiently detected even when they are not among high-frequency topics?
- **RQ (11) Sentiment Classifier for Short Texts**: What is the best choice so far for classifying sentiment on social media when domain of texts is unknown?
- **RQ (12) Sentiment Reason Mining Dashboard**: How to ensure that Human Machine Interface displays Reason Candidates in a comprehensible form?

1.3 Research Objectives

The main objective of this thesis is to develop an enhanced Emerging Topic Detection Model that can identify hot topics when two large sets of timestamped documents are compared. The model shall be used to build a Sentiment Reason Mining framework that can interpret sentiment changes on Twitter for any given target by the user. It shall automatically extract main reasons behind each major sentiment variation without the need of user intervention. This thesis also aims to examine existing Sentiment Reason Mining methods to identify main research gaps. To the best of our knowledge, this work offers the first attempt to compare and assess existing methods through qualitative analysis and experiments on real-life data.

1.4 Summary of Contributions

Most important contribution of this thesis is making it possible to interpret sentiment changes on social media automatically and accurately through a novel model, called Filtered-LDA. A user would just write a query on a social media target and set an alarm threshold for high positive/negative sentiment variations, then the system automatically tracks actual changes online and extracts their reason candidates. The new model can be used for multiple applications, wherein emerging topics shall be detected in a large set of timestamped documents.

Other main contributions can be summarized as follows:

- Create transparency on existing main Sentiment Reason Mining methods. To the best of our knowledge, this thesis presents the first survey that attempts to investigate existing methods of both Sentiment Reasoning and Sentiment Variations' Reasoning tasks.
- Find main research gaps in the related studies by applying them on real-life data.
- Compare state-of-the-art Sentiment Analysis classifiers on social media when Machine Learning training domains are different from testing domains.
- Develop mechanisms to enhance performance of Topic Models for tracking topics over time. Multiple frameworks are proposed to implement these mechanisms.

1.5 Thesis Outline

This thesis starts with a qualitative review of existing methods, followed by testing main Sentiment Variations' Reason Mining methods. Afterwards, a model is designed to tackle existing challenges. To select the right Sentiment Analysis tool for the new model, a comparison is then conducted for main classifiers. Subsequently, the proposed new technique is applied on a real-life large dataset.

Chapter 2 explains the importance of the Sentiment Reason Mining task and its role in future Sentiment Analysis research directions. Then it conducts a survey on existing methods, and it clarifies the difference between Sentiment Reasoning methods and Sentiment Variations' Reasoning methods. A brief description for each method is presented along with examples on related literature.

Chapter 3 selects two Twitter datasets for testing main Sentiment Variations' Reasoning methods. Multiple experiments are then conducted to examine existing hypothesis and find main research gaps. The chapter is concluded with discussing the experimental results.

Chapter 4 selects the Emerging Topic Detection approach to handle the Sentiment Variations' Reasoning task. It starts with reviewing base models, followed by testing them on a real-life dataset. Subsequently, a new model is proposed to overcome existing limitations. Finally, the chapter proposes two different frameworks to implement the proposed new model.

Chapter 5 focuses on the Sentiment Analysis function which shall be used for the proposed Sentiment Reasoning Model. It applies state-of-the-art Sentiment Analysis tools and classifiers on two different Twitter datasets that belong to different domains. The chapter then concludes the top performing classifier that shall be used for the new model.

Chapter 6 presents a framework for Sentiment Reason Mining based on the developed model for Emerging Topic Detection. Then it tests the performance of the proposed framework on a Ground Truth dataset, which is extracted from a real-life large Twitter dataset. Subsequently, a similar test is conducted on the complete large dataset to automatically extract Reason Candidates for sentiment variations. Finally, the chapter proposes a simple method for presenting the output of the proposed framework so that it can be easily interpreted by users.

Chapter 7 summarizes main conclusions of the thesis, then answers research questions, which are raised in Section 1.2.

1.6 Publications Overview

This thesis is publication-based as its core material is derived from the manuscripts that have been submitted to a peer reviewed academic journal. The following 1st and 3rd papers had been accepted by the IEEE Access Journal upon completing the research and coding works by the first author, whereas the 2nd paper is still under review by the MDPI AI Journal:

1) Alattar, F. & Shaalan, K., 2021. A Survey on Opinion Reason Mining and Interpreting Sentiment Variations. IEEE Access Journal, Volume 9, pp. 39636-39655.
2) Alattar, F. & Shaalan, K., 2021. Emerging Research Topic Detection using Filtered-LDA. MDPI Artificial Intelligence Journal, Volume XX, pp. XX-XX.
3) Alattar, F. & Shaalan, K., 2021. Using Artificial Intelligence to Understand What Causes Sentiment Changes on Social Media. IEEE Access Journal, Volume 9, pp. 61756-61767.

Chapters 2 and 3 are derived from the first paper, whereas Chapter 4 is derived from the second paper. Finally, the material of the third paper is used for Chapters 5 and 6.

Chapter 2 : Literature Review*

2.1 Introduction

Sentiment Analysis 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 of D'Andreaet al. (2015) for more examples.

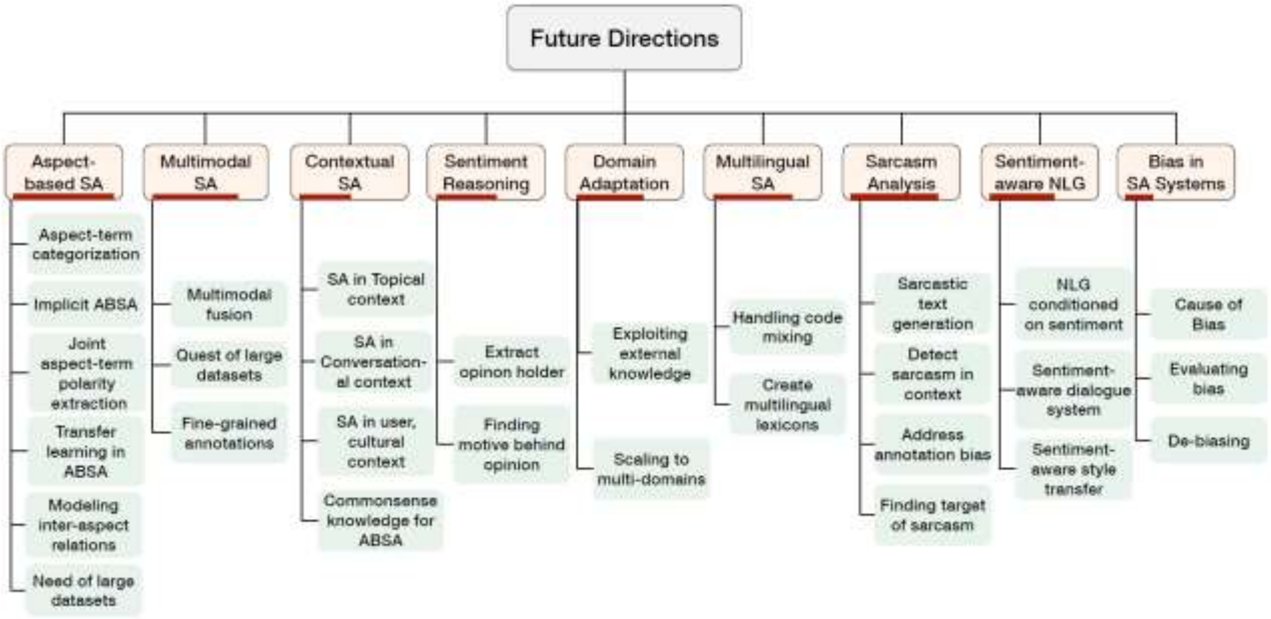

Future Directions
Aspect-based SA
Aspect-term categorization
Implicit ABSA
Joint aspect-term polarity extraction
Transfer learning in ABSA
Modeling inter-aspect relations
Need of large datasets
Multimodal SA
Multimodal fusion
Quest of large datasets
Fine-grained annotations
Contextual SA
SA in Topical context
SA in Conversation-al context
SA in user, cultural context
Commonsense knowledge for ABSA
Sentiment Reasoning
Extract opinon holder
Finding motive behind opinion
Domain Adaptation
Exploiting external knowledge
Scaling to multi-domains
Multilingual SA
Handling code mixing
Create multilingual lexicons
Sarcasm Analysis
Sarcastic text generation
Detect sarcasm in context
Address annotation bias
Finding target of sarcasm
Sentiment-aware NLG
NLG conditioned on sentiment
Sentiment-aware dialogue system
Sentiment-aware style transfer
Bias in SA Systems
Cause of Bias
Evaluating bias
De-biasing

Figure 1: Future Directions of Sentiment Analysis Research (Poria et al., 2020).

* This chapter is derived from the article "Alattar, F. & Shaalan, K., 2021. A Survey on Opinion Reason Mining and Interpreting Sentiment Variations. IEEE Access, Volume 9, pp. 39636-39655".

Most of Sentiment Analysis studies - like (Panget al., 2002), (Baneaet al., 2008), (Wilsonet al., 2009), and (Cambriaet al., 2013) - 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 2.2 & 2.3.

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. (2020) 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 focus of this survey is the Sentiment Variations' Reasoning problem and its application on Twitter. Our objective is to qualitatively examine various methods for solving this problem, then we shall analyze them in Chapter 3 through real-life datasets to discover research 'Empirical Gaps' (Miles, 2017) and "Evaluation Voids" (Muller-Bloch & Kranz, 2015). After exploring these methods, main hypothesis used in relevant studies are discussed. Sections 2.2 and 2.3 investigate main Reason Mining approaches in the field of Sentiment Analysis by selecting representative articles for each approach. Fig. 2 summarizes these approaches.

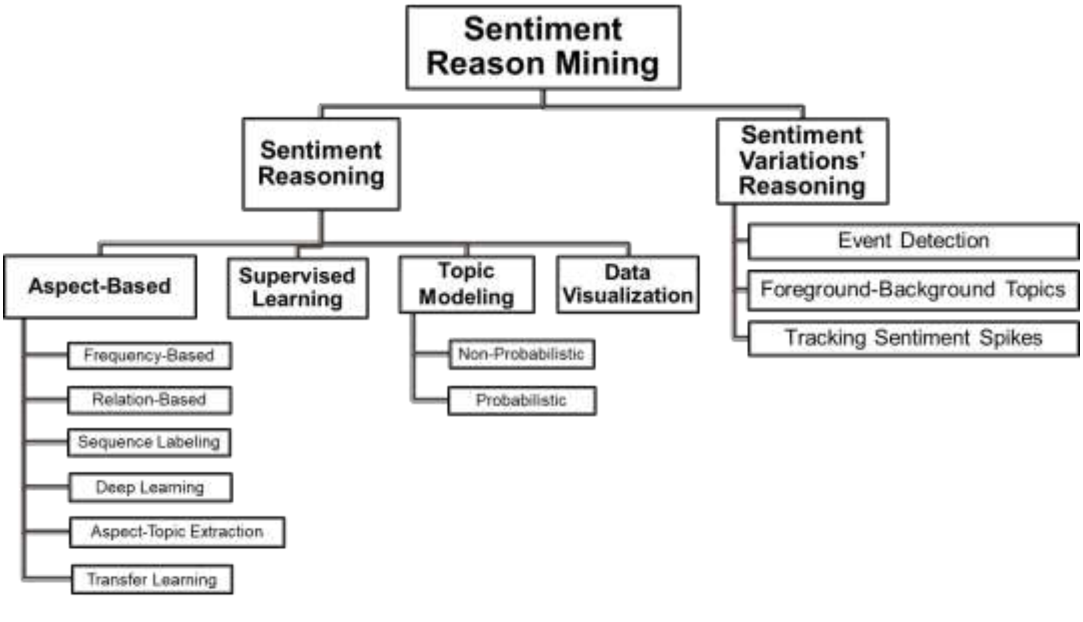


Figure 2: Approaches of Reason Mining for Sentiment Analysis.

2.2 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. 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.

2.2.1 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 (Pang & Lee, 2008). 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 (Liu, 2010). 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 (Zhang & Liu, 2014). 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 (Wang et al., 2015).

2.2.1.1 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 & Liu (2004) 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 & Liu (2004) 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 & Etzioni (2005) enhanced the approach of Hu & Liu (2004) 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 (Hu & Liu, 2004).

To enhance the Frequency-Based method further, Scaffidi et al. (2007) compared the frequency of detected nouns and noun-phrases with their frequency in English-Corpus. Blair-Goldensohn et al. (2008) also aimed to enhance the method by considering the nouns which are included in subjective sentences.

O'Connor et al. (2010) 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 by Moghaddam & Ester (2010) through 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 (Ishaq, Asghar & Gillani, 2020).

2.2.1.2 Relation-Based

The Relation-Based approach, also known as Rule-Based and Syntax-Based (Rizvana et al., 2018), 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. (2006) 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. (2009) enhanced the Relation-Based method by using a phrase-dependency-parser to extract aspects which are noun-phrases or verb-phrases. Qui et al. (2011) employed a double-propagation technique to identify relationships between opinion words and aspects.

Jiang et al. (2011) 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 & Ghotkar (2016) argued that this approach may produce many terms which are not real aspects. It is also confirmed by Syamala & Nalini (2019) that Relation-Based approach may extract irrelevant features.

2.2.1.3 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. Jin et al. (2009) used HMM in a framework that integrates linguistic elements like Part-Of-Speech (POS) into a learning process to predict patterns and correlations between tags. Choi & Cardie (2010) used a CRF model to detect boundaries of sentiment phrases and identify both polarity and intensity of these phrases.

A Hierarchical Sequence Labeling Model (HSLM) was developed by Chen et al. (2020). 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 by Syamala & Nalini (2019) 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. Wang & Ren (2020) used Sequence-to-Sequence (Seq2Seq) learning to reduce the impact of this limitation. Seq2Seq considers relations between the opinion polarity of features in feature-level opinion classification method.

2.2.1.4 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. Liu et al. (2015) used RNN with Word Embedding to develop a discriminative model for Aspect-Extraction. Similarly, RNN is used by Jebbara & Cimiano (2016) for the Two-Step Aspect-Extraction method. RNN is also applied by Yang et al. (2018) for the Financial domain on tweets and news headlines. CNN is used by both Xu et al. (2016) and Poria et al. (2016) to carry out aspect-extraction as a multi-label classification problem, whereas it is combined with RNN by Dhanush et al. (2016) to handle Aspect-Extraction and Sentiment Analysis tasks.

The mechanism of Attention Models was introduced to enhance the performance of Deep Learning methods. Zhang & Lu (2019) developed a Multi-Attention Network to handle the Aspect-Extraction task.

It is indicated by Saraiva et al. (2020) 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 by Zohuri & Moghaddam (2020) 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.

2.2.1.5 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 (Sharma, 2020).

To illustrate, the Topic Model of Latent Dirichlet Allocation (LDA) (Blei et al., 2003) 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” (Sharma, 2020). 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 (Lu et al., 2011). LSA is used by Lu et al. (2009) to create a feature summary with a rating for each feature. Similarly, it is employed by Ortega et al. (2014) to build a domain-dependent aspect-extraction sentiment analysis framework. LDA is used by Brody & Elhadad (2010) to handle the aspect-extraction task.

Titov & McDonald (2008) 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.

Li et al. (2010) 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 & Maalej (2014) 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 (Thelwall et al., 2010), to carry out Sentiment Analysis. Finally, they used LDA Topic Model to group fine-grained aspects into more expressive high-level aspects.

ORMFW (Khalid & Khan, 2018) 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.

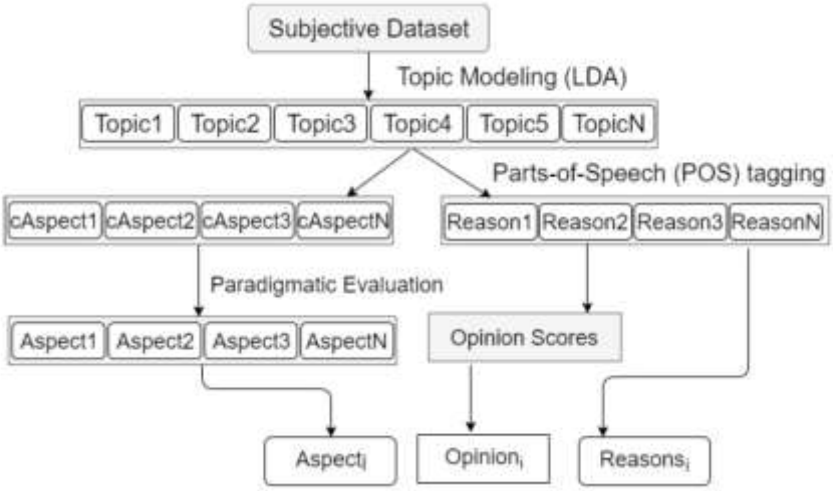


Figure 3: ORMFW, Opinion Reason Mining Framework (Khalid & Khan, 2018)

Later, Khalid et al. (2018) 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. (2018) 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 (Chen et al., 2018).

2.2.1.6 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 (Zhuang et al., 2021). Fig. 4 shows a visual example to simplify the concept of Transfer Learning. A person who is familiar with bicycles can transfer this knowledge to the motorcycle's domain.

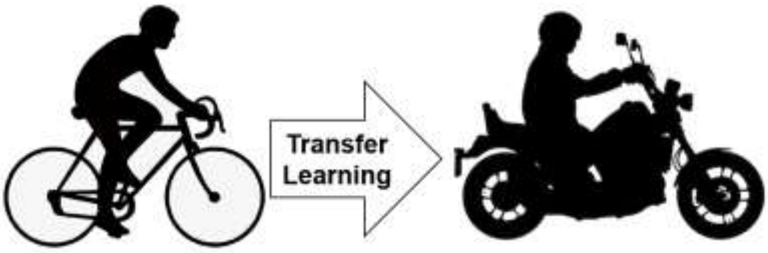


Figure 4: Perceptive example about Transfer Learning

Tao & Fang (2020) 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.

2.2.2 Supervised Learning

As indicated by Poria et al. (2020), 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 & Hovy (2006) 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.

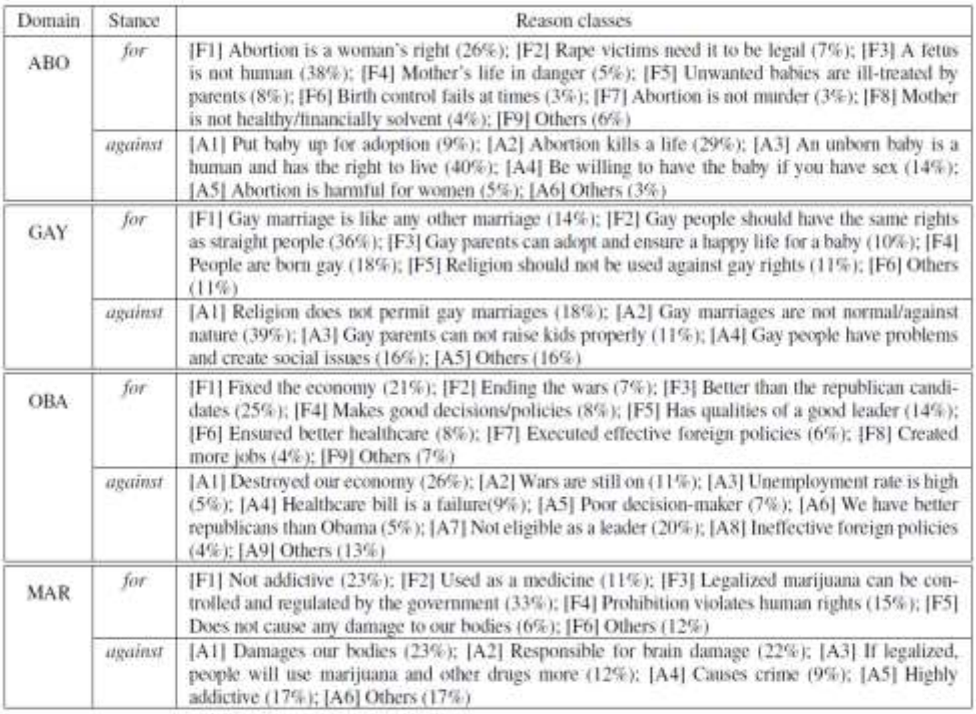
Li et al. (2006) used Support Vector Machines (SVM) and Naive Bayes (NB) to learn perspective of an opinionated text on both document-level and sentence-level basis. To achieve this task, they created a corpus with label perspective on a document-level, however they could not label it on sentence-level, 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. (2007) 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 by Yessenalina et al. (2010) through labeling the rationales automatically, rather than using human annotation.

Persing & Ng (2009) 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.

Boltuzic & Snajder (2014) prepared Corpus of Online User Comments with Arguments (ComArg, 2014), 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 Boltuzic & Snajder (2014), Hasan & Ng (2014) 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.



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%)

Table 1: Reason Classes for Debate Forums (Hasan & Ng, 2014)

The Supervised Learning approach requires data annotation work, which can be achieved either automatically (Kim & Hovy, 2006) for product reviews or manually (Hasan & Ng, 2014) 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.

2.2.3 Topic Modeling

Utilization of Topic Models for extracting Aspects from text has been addressed in subsection 2.2.1.5. 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. (2014) 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 framework, e.g., (Eguchi & Lavrenko, 2006).
2) Carry out each task separately for the same set of documents, e.g., (Hurst & Nigam, 2004).
3) Consider one of the tasks as a prior to the other, e.g., (Eguchi & Shah, 2006).

In general, Topic Models use statistical methods to discover topics that appear in a set of either short or long texts. Some of these models are non-probabilistic, like the Latent Semantic Indexing (LSI) - also known as Latent Semantic Analysis (LSA) (Deerwester et al., 1990) and the Non-Negative Matrix Factorization (NMF) which became popular when (Lee & Seung, 1999) employed it for Topic Modeling.

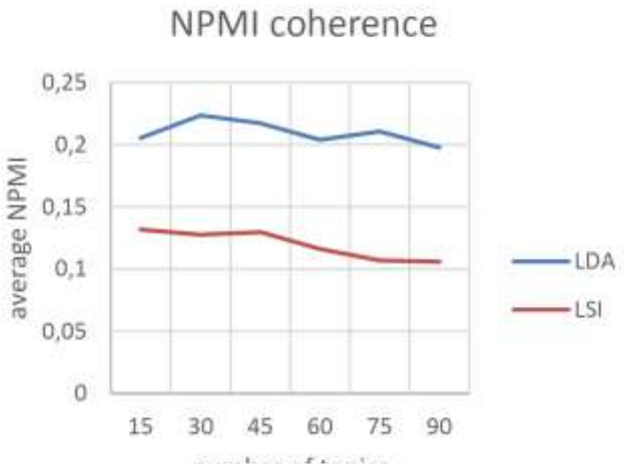


Figure 5: LDA vs LSI Topic coherence for Reuters dataset (Smatana et al., 2019)

The second category of Topic Models that gained popularity during the last two decades is the probabilistic topic modeling. Hofmann (1999) transformed the LSI method into a probabilistic model called Probabilistic Latent Semantic Analysis (PLSA) or Probabilistic Latent Semantic Indexing (PLSI). Later, Blei et al. (2003) 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 (Blei & McAuliffe, 2010).

When LSA and LDA methods are compared, LDA can extract more coherent topics (Chiru et al., 2014). Smatana et al. (2019) 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 (Cvitanic et al., 2016). Furthermore, it is likely that 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 (Rehurek, 2021) library for Python, tomotopy (Bab2min, 2021b) library provides Python programmers with a wide range of Topic Modeling algorithms, which include:

- Latent Dirichlet Allocation (Blei et al., 2003)
- Hierarchical LDA (Blei et al., 2004)
- Hierarchical Dirichlet Process (Teh et al., 2004)
- Correlated Topic Model (Blei & Lafferty, 2005)
- Pachinko Allocation (Li & McCallum, 2006)
- Dynamic Topic Model (Blei & Lafferty, 2006)

- Hierarchical Pachinko Allocation (Mimno et al., 2007)
- Multi Grain LDA (Brody & Elhadad, 2010)
- Labeled LDA (Ramage et al., 2009)
- Supervised LDA (Blei & McAuliffe, 2010).
- Partially Labeled LDA (Ramage et al., 2011)
- Dirichlet Multinomial Regression (Mimno & McCallum, 2008)
- Generalized Dirichlet Multinomial Regression (Lee & Song, 2020)

Mei et al. (2007) 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 & Lee (2008) 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. (2015) 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.

2.2.4 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 (Tan et al., 2014).

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

TimeMines (Swan & Jensen, 2000) 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 (Havre et al., 2002) is a Topic Visualization software that was developed to track topic evolution within a large set of documents.

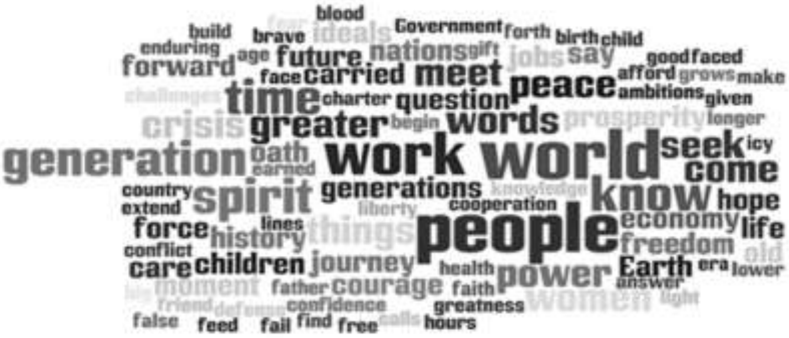


Figure 6: Word-cloud representation (Hofmann & Chisholm, 2016)

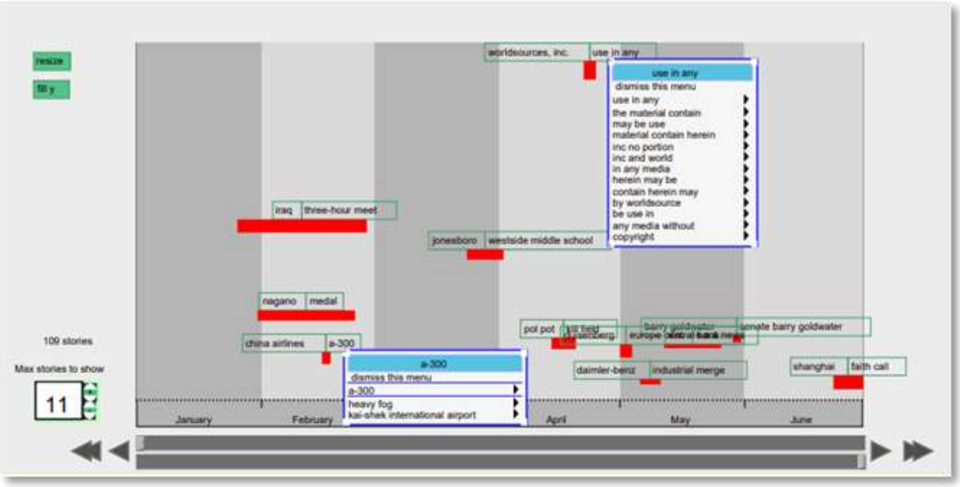


Figure 7: TimeMines Topic Visualization (Swan & Jensen, 2000)

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., 2002, 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 (Van Eck & Waltman, 2010) 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 during the selected period. Fig. 9 shows how VOSviewer assigns colors to topic terms based on their frequency and average years of publications.

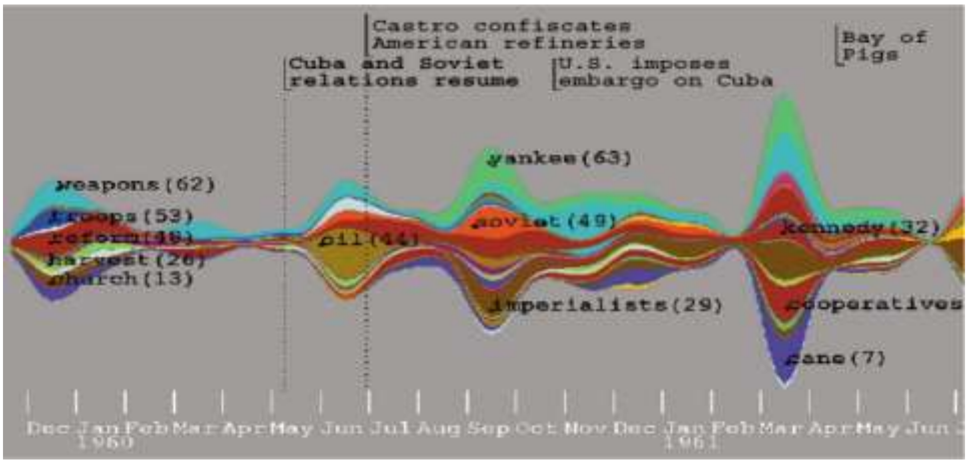
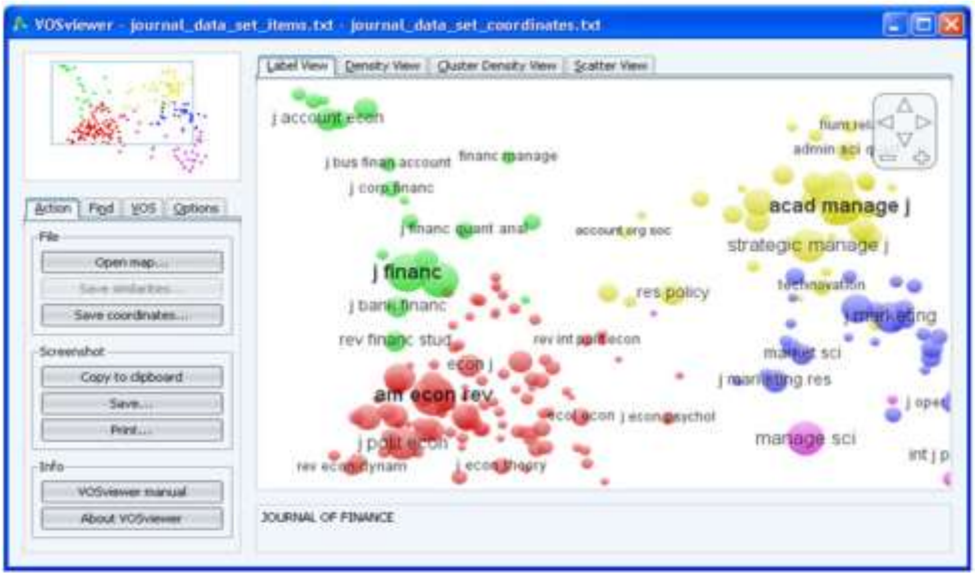


Figure 8: ThemeRiver Topic Visualization (Havre et al., 2002)



JOURNAL OF FINANCE

Figure 9: VOSviewer cluster density view of a journal map (Van Eck & Waltman, 2010)

With the help of VOSviewer software, Shen & Wang (2020) 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 (Vorontsov et al., 2015) 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 (2016) used VisARTM to obtain hierarchic topic visualization using polygons which show short description of documents along with the main topic labels as shown in Fig. 11. Such representation ensures better understanding of each topic without the need of skimming

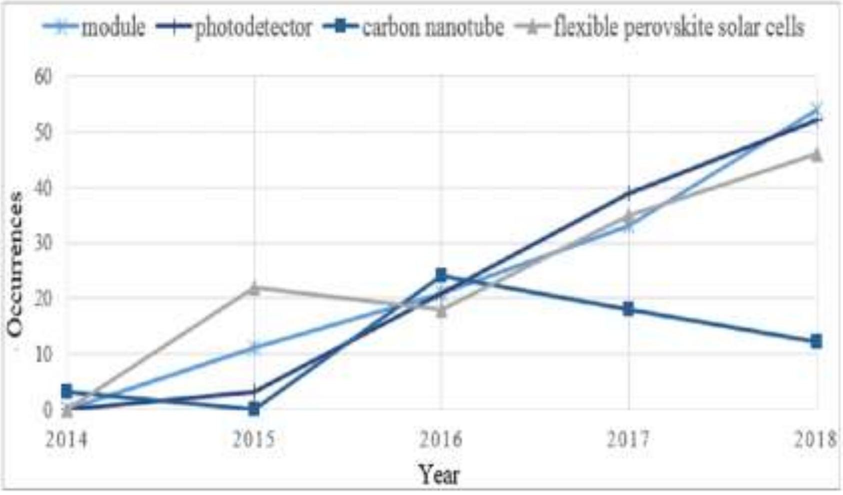


Figure 10: Variation in PSC term occurrence (Shen & Wang, 2020)

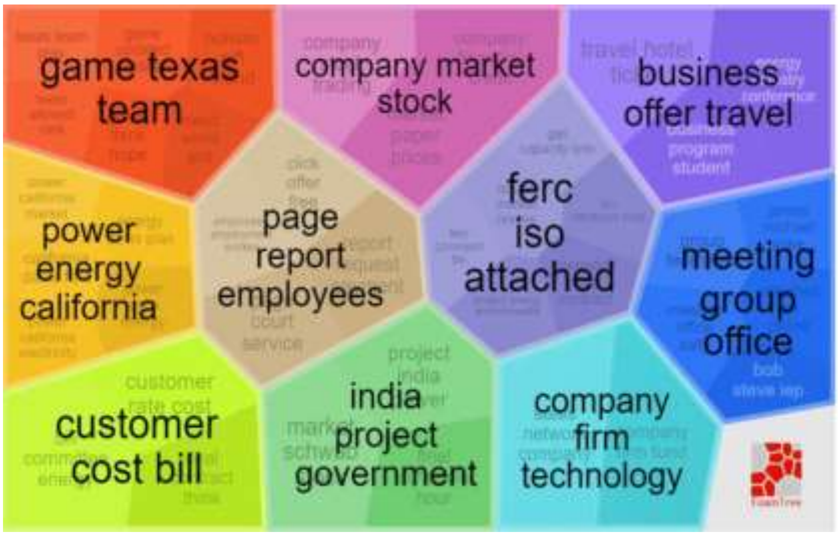


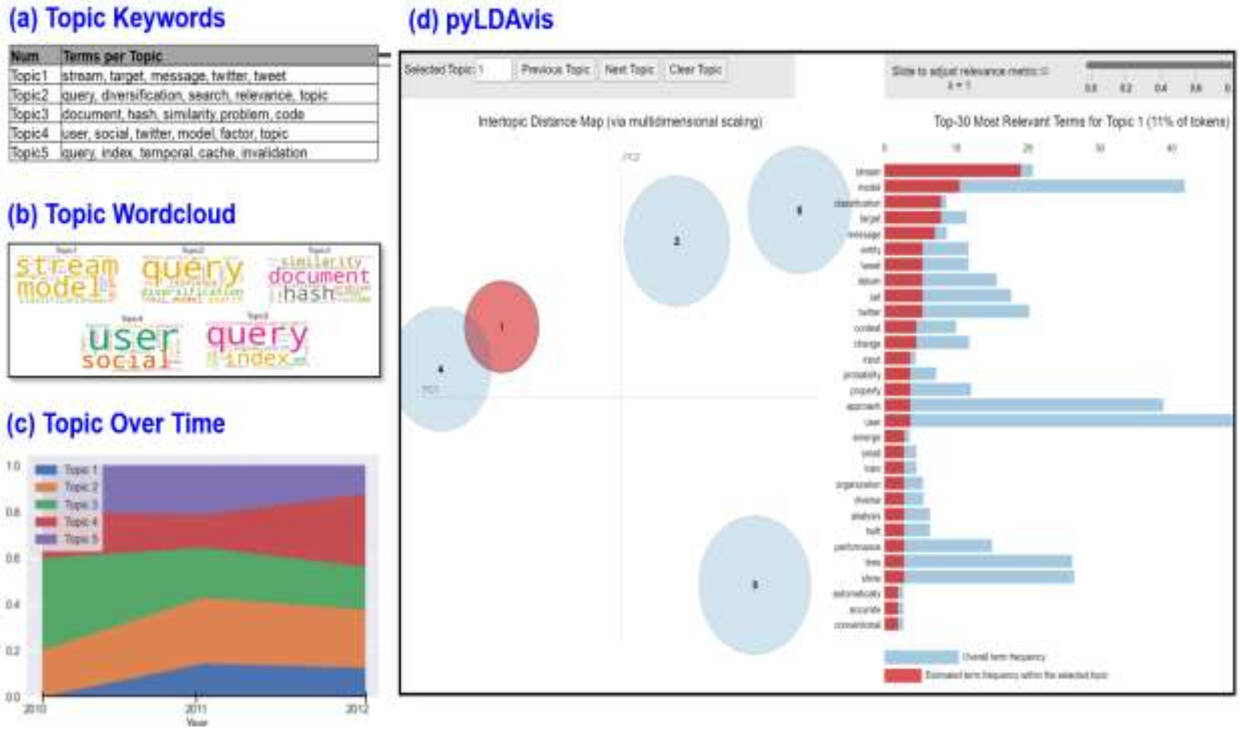
Figure 11: VisARTM document representation (Fedoriaka, 2016)

through actual documents. However, VisARTM suffers a limitation when handling large sets of text because it is a web-based tool.

Smatana et al. (2019) 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, we applied LDA topic modeling 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 (Sievert & Shirley, 2014) which was introduced for the first time in year 2014. 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.



(a) Topic Keywords

Num	Terms per Topic
Topic1	stream, target, message, twitter, tweet
Topic2	query, diversification, search, relevance, topic
Topic3	document, hash, similarity problem, code
Topic4	user, social, twitter, model, factor, topic
Topic5	query, index, temporal, cache, invalidation

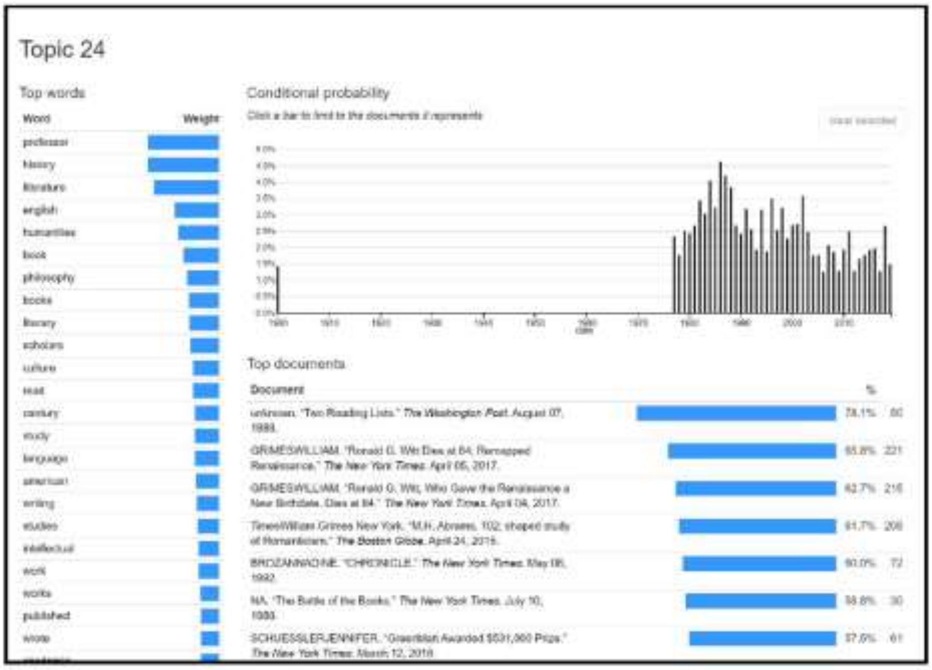
(b) Topic Wordcloud

(c) Topic Over Time

(d) pyLDAvis

Figure 12: Python Topic Visualization for SIGIR 2010-2012 dataset

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.



Topic 24

Top words

Word	Weight
professor	
history	
literature	
english	
humanities	
book	
philosophy	
books	
literary	
scholars	
culture	
read	
century	
study	
language	
american	
writing	
studies	
intellectual	
work	
works	
published	
wrote	

Conditional probability

Click a bar to limit to the documents it represents

Top documents

Document	%	
unknown. "Two Reading Lists." *The Washington Post*. August 07, 1988.	78.1%	60
GRIMESWILLIAM. "Ronald G. Witt Dies at 84; Remapped Renaissance." *The New York Times*. April 05, 2017.	65.8%	221
GRIMESWILLIAM. "Ronald G. Witt, Who Gave the Renaissance a New Birthdate, Dies at 84." *The New York Times*. April 04, 2017.	62.7%	215
TimesWilliam Grimes New York. "M.H. Abrams, 102; shaped study of Romanticism." *The Boston Globe*. April 24, 2015.	61.7%	206
BROZANNADINE. "CHRONICLE." *The New York Times*. May 08, 1992.	60.0%	72
NA. "The Battle of the Books." *The New York Times*. July 10, 1988.	58.8%	30
SCHUESSLERJENNIFER. "Guardian Awarded $531,000 Prize." *The New York Times*. March 12, 2018.	57.6%	61

Figure 13: dfr-browser topic representations (Goldstone et al., 2014)

Dfr-browsers (Goldstone et al., 2014) is a free web-based tool for browsing topics from a set of documents. Fig. 13 shows sample of 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.

2.3 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.

2.3.1 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 (Tan et al., 2014).

Outside the Sentiment Reasoning field, many event detection techniques were developed to capture spiky topics in an online data stream. Leskovec et al. (2009) 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 by Sakaki et al. (2010) 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 by Chakrabarti & Punera (2011). 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.

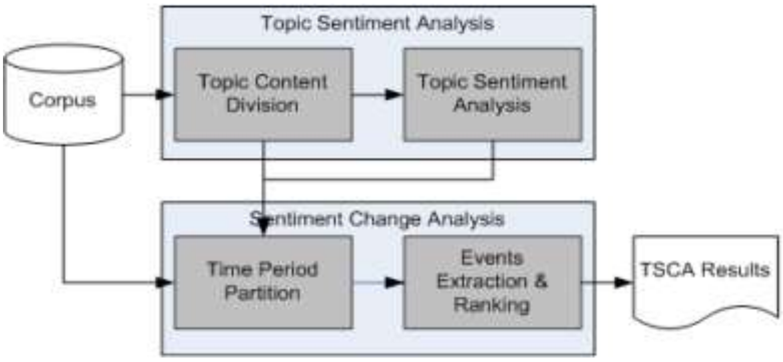


Figure 14: Topic Sentiment Change Analysis Framework (Jiang et al., 2011)

Weng & Lee (2011) 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 Event-Detection with Clustering of Wavelet-based Signals (EDCoW), 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. (2011) 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 (Mei et al., 2007), 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. (2011) 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 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

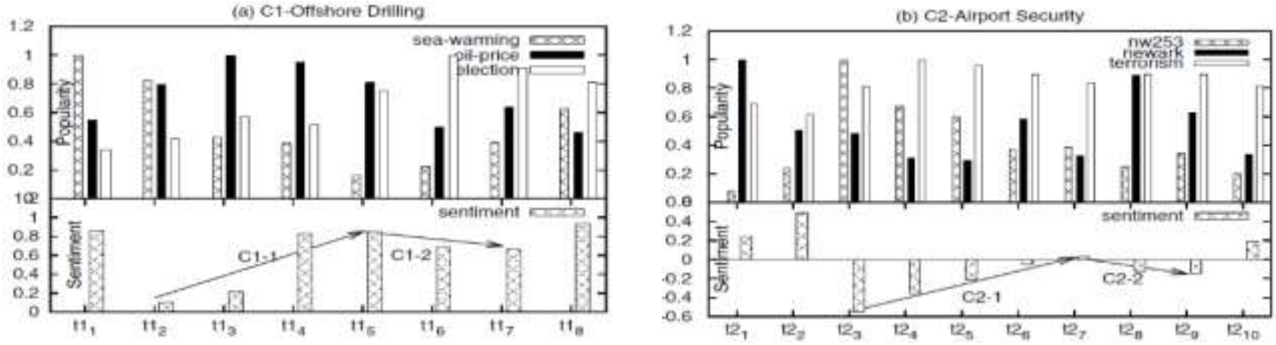


Figure 15: Sentiment Variation Reasoning using PLSA and TTP (Jiang et al., 2011)

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. (2011) 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 by Jiang et al. (2011).

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. Meng et al. (2012) developed a method 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 (Abdelhaq et al., 2013) 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. (2015) 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.

Chen et al. (2017) used a Supervised Learning approach 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. (2020) 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. (2018) 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 (Hettiarachchi et al., 2020) 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. Hettiarachchi et al. (2020) also tried to use other word embedding methods like BERT, however it was realized that such advanced methods need relatively long learning times.

2.3.2 Foreground-Background Topis

This method was introduced by Tan et al. (2014) 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 variation – by removing from the Foreground all old topics that appeared in the Background period, which represents the duration before the Foreground period. For

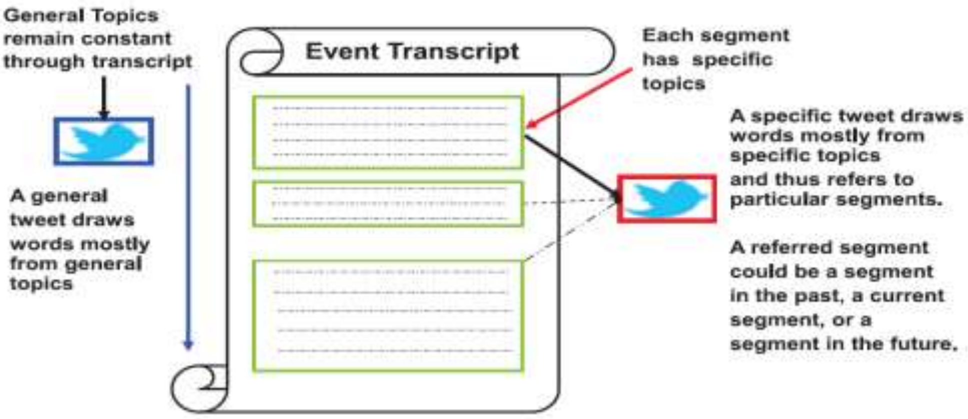

General Topics remain constant through transcript
Event Transcript
Each segment has specific topics
A specific tweet draws words mostly from specific topics and thus refers to particular segments.
A general tweet draws words mostly from general topics
A referred segment could be a segment in the past, a current segment, or a segment in the future.

Figure 16: Conceptual ET-LDA model (Hu et al., 2012)

extracting the tweets' sentiment, they combined both SentiStrength and TwitterSentiment tools in a hybrid approach.

The FB-LDA Model was inspired by the article of Hu et al. (2012) who built an event detection and segmentation system. For an event that appears inside a large-scale group of tweets, Hu et al. (2012) 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. (2012) 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. (2014) 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. They also used 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.

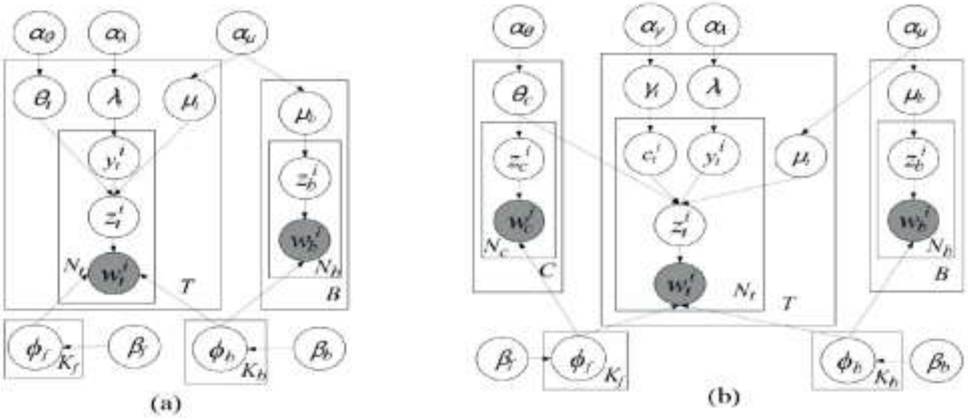


Figure 17: (a) FB-LDA and (b) RCB-LDA models (Tan et al., 2014)

Although the developers of FB-LDA did not focus on the Sentiment Analysis task itself, their Reason Mining work was followed by series of projects through other researchers – e.g. (Ingule & Chhajed, 2014), (Poonam & Kinikar, 2014), (Patil, Sedamkar & Gupta, 2015), (Bhalerao & Dange, 2015), (Urega & Devapriya, 2015), (Jamadar et al., 2016), (Kamini & Ezhillarasi, 2015), (Avachar et al., 2016), (Patil & Kulkarni, 2018), (Jeevitha, 2016), (Manikandan & Kalpana, 2016), and (Admane et al., 2016) - who tried to use better Sentiment Analysis classifiers for the FB-LDA Model. To illustrate, Fig. 18 shows a Reason Mining framework (Patil, Sedamkar & Gupta, 2015) that utilizes FB-LDA and RCB-LDA models.

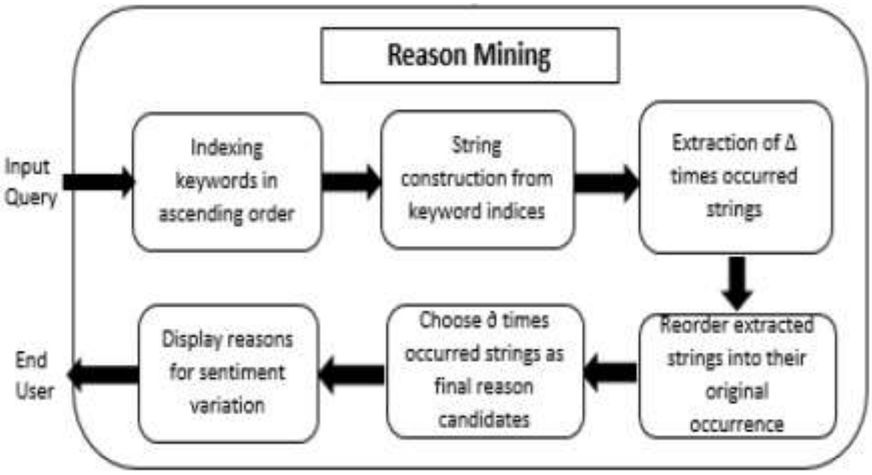

Reason Mining
Input Query
Indexing keywords in ascending order
String construction from keyword indices
Extraction of Δ times occurred strings
Reorder extracted strings into their original occurrence
Choose ẟ times occurred strings as final reason candidates
Display reasons for sentiment variation
End User

Figure 18: Reason Mining Framework using FB-LDA (Patil, Sedamkar & Gupta, 2015)

2.3.3 Tracking Sentiment Spikes

This method was introduced by Giachanou, Mele, & Crestani (2016) who (i) extracted tweets' sentiment using SentiStrength (Thelwall et al., 2010) 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 (Giachanou & Crestani, 2016b). 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-a 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, Mele, & Crestani (2016) 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 two probability distributions, and it was employed by Giachanou & Crestani, (2016b) to rank the discovered topics inside the sentiment spike. Fig. 19-b demonstrates some topic trends inside a sentiment spike.

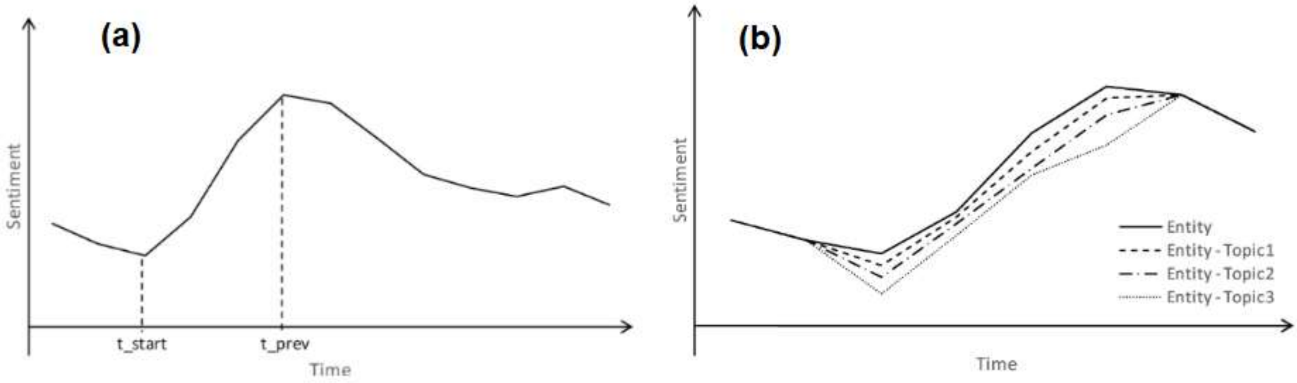


Figure 19 (a) Sentiment spike (b) Topics inside a sentiment spike (Giachanou, Mele, & Crestani, 2016)

2.4 Summary

This chapter presented the main approaches that addressed the Sentiment Reason Mining task. Representative literatures were reviewed to understand each approach, with focus on the Sentiment Variations’ Reasoning methods, namely the Event Detection, FB-LDA, and Tracking Sentiment Spikes.

Sentiment Reasoning through Aspect-Based methods is useful, especially in products and services domains when features of the target can be identified in advance. Supervised-Learning methods are also useful for limited number of applications where a Reason-Annotated Corpus can be created, however the performance of such methods is relatively low, though they demand a lot of resources for data annotation.

Topic Models play main role in Sentiment Reasoning, especially when they are combined with proper visualization dashboards. However, fine-tuning these models to achieve accurate results faces genuine challenges, like identifying the number of topics in advance and selecting the right hyperparameters’ values. Next chapter will examine main approaches of Sentiment Variations’ Reasoning using real-life datasets.

Chapter 3 : Background Work on Sentiment Variations' Reasoning Methods*

To address main research questions RQ (1) to RQ (9), which are listed in Section 1.2, two Twitter datasets are analyzed: First one is a pre-annotated 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.

3.1 Experimental Setup

We used a Dell Inspiron 7370 laptop with Intel$^{(R)}$ Core$^{(TM)}$ i7-8550U CPU @ 1.99 GHz Processor, Installed RAM is 16.0 GB (15.8 GB usable), Windows 10 Home version 20H2 64-bit operating system. Python version 3.8.3 is used with Jupyter Notebook server version 6.1.4.

To apply Topic Modeling on the "Apple Tweets" dataset, we imported LdaMallet from Gensim model's wrapper with default hyperparameter settings, and we turned on the "optimize interval" option, which optimizes hyperparameters every 10 training iterations.

3.2 Datasets

The first dataset is the famous US Airlines Twitter Dataset (Crowdower, 2015). 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 US Airlines Twitter Dataset is considered a relatively small-size dataset, however we selected it because it includes the Neutral Sentiment labels. Unfortunately, the available large-size Twitter datasets do not include the Neutral Sentiment labels.

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 (Kwak et al., 2010). 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

* This chapter is derived from the article "Alattar, F. & Shaalan, K., 2021. A Survey on Opinion Reason Mining and Interpreting Sentiment Variations. IEEE Access, Volume 9, pp. 39636-39655".

polarity of each of the remaining 7,079 tweets, which include 1,733 negative, 2,873 neutral, and 2,473 positive tweets.

We selected the large-size Stanford Twitter Dataset so that we can compare our results later with Tan et al. (2014) who had chosen the same dataset to test their FB-LDA model.

3.3 Explicit Reasons

By analyzing the reasons/categories of negative tweets in the US Airlines dataset – see Fig. 20 –it is noticed that 87% of the negative tweets explicitly indicate the reason behind negative 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 tweets were analyzed, then we manually labeled the concluded reason behind sentiment. Fig. 21 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.

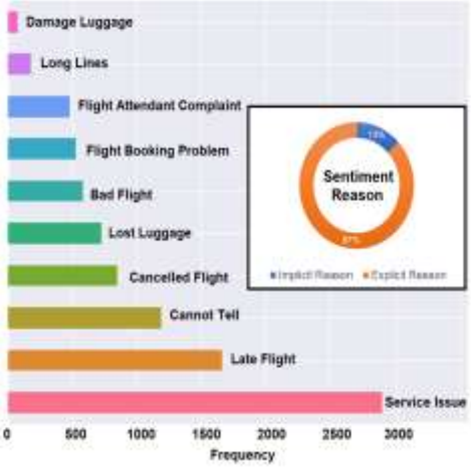


Figure 20: Sentiment Reasons in US Airlines Negative Tweets

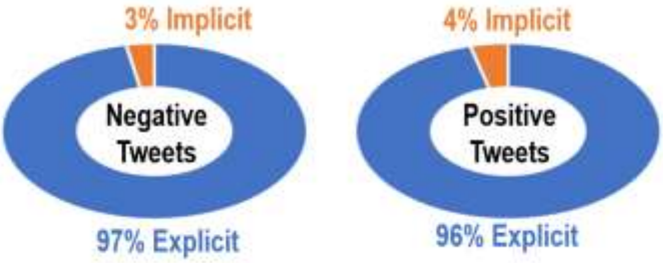


Figure 21: Implicit vs Explicit Sentiment Reasons in ‘Apple” tweets

Hence, 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.

3.4 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. 22 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.

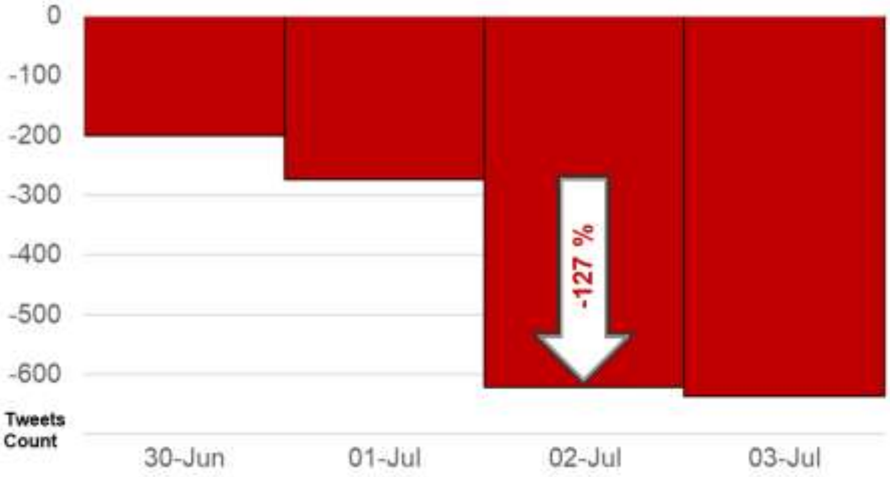
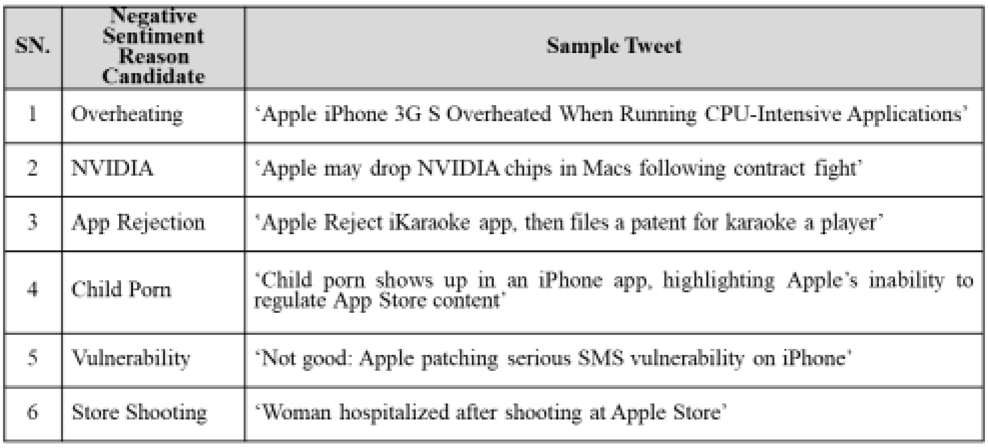


Figure 22: Number of Tweets inside Apple's negative sentiment spike

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. 23.

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



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'

Table 2: Reason Candidates for Apple's Negative Sentiment

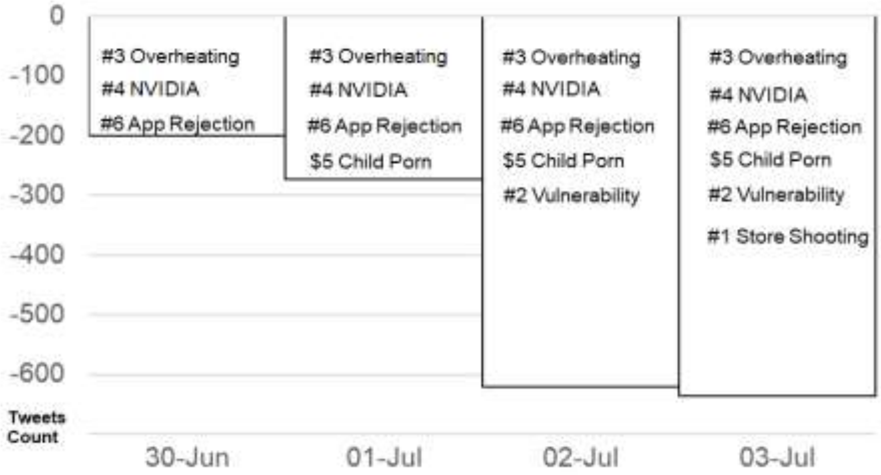
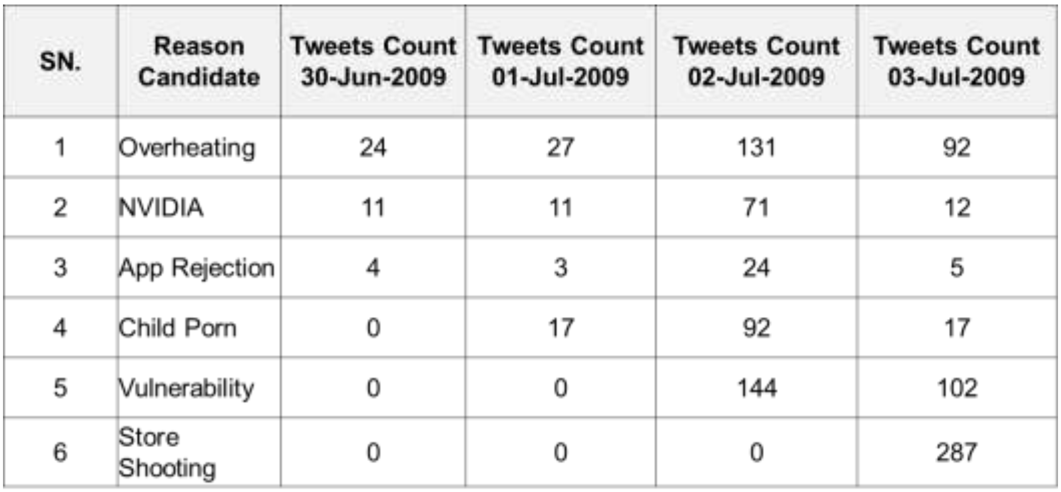


Figure 23: Topics Evolution inside Apple's negative sentiment spike

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, Aspect-Based method cannot always capture reasons of sentiment for products and services domains.

3.5 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, main reason for a sentiment spike does not always have the highest topic frequency.



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

Table 3: Count of Topic Tweets for Apple's Negative Sentiment

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 "Vulnerability" Emerging Topic is the main reason for the spike as it did not exist before 02-Jul-2009.

3.6 Events

As shown in Fig. 22, there is no major change in overall "Apple Tweets" sentiment level on 01-Jul-2009, therefore Tracking Sentiment Spikes' method would not capture this 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 indicates that the Event Detection method cannot address the Reason Mining task for the sentiment level change on 01-Jul-2009.

3.7 Emerging Topics

In the above subsections, 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 sense to employ the Foreground-Background Topics method for the Reason Mining task.

3.8 Topic Visualization

To examine the role of Topic Visualization in the Sentiment Reasoning task, Fig. 24 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.

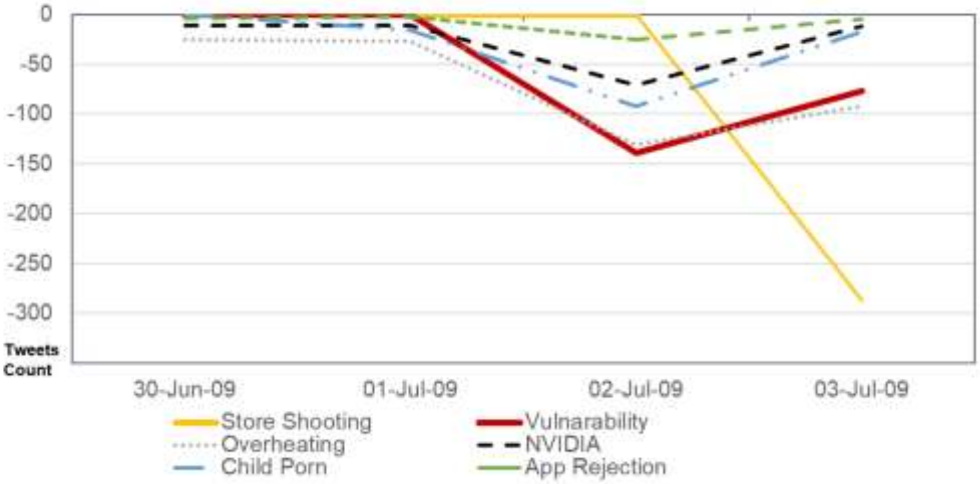


Figure 24: Topics Evolution inside Apple’s negative sentiment spike

It is therefore evident that Topic Visualization does genuinely enhance our understanding of the Sentiment Reasoning.

3.9 Sentiment Spikes

For “Apple Tweets”, Fig. 25 is drawn to show the trends of “Overall Negative Sentiment”, the “Vulnerability” Topic, and the “Store Shooting” Topic. This figure represents 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, sentiment variation reason does not always cause a spike in the overall sentiment level.

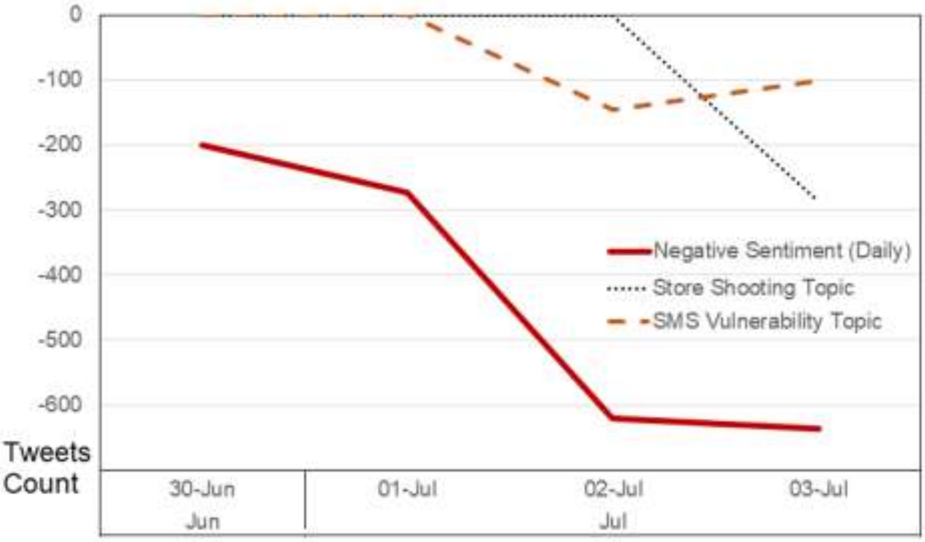


Figure 25: Overall Negative Sentiment vs Top Reason Candidates

3.10 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. 26. 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.

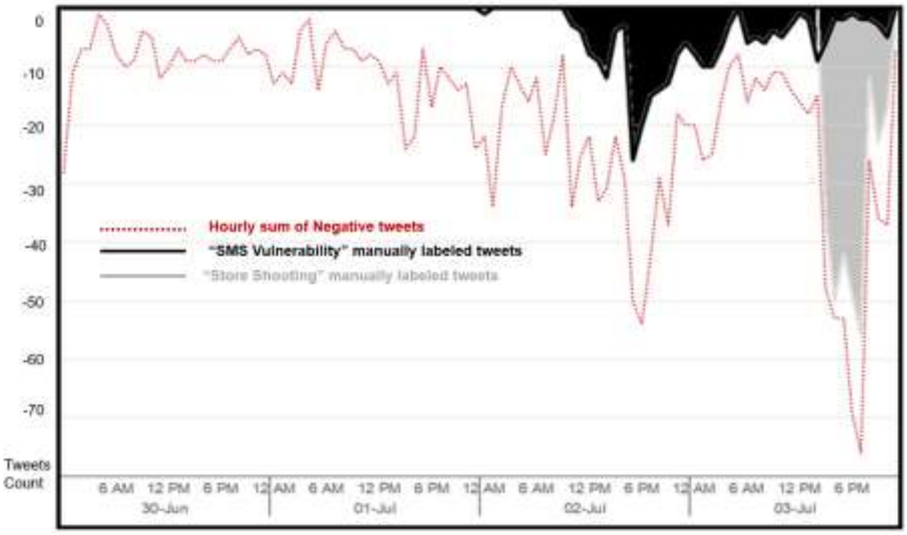


Figure 26: Negative Sentiment Level vs Manually Labeled Topics

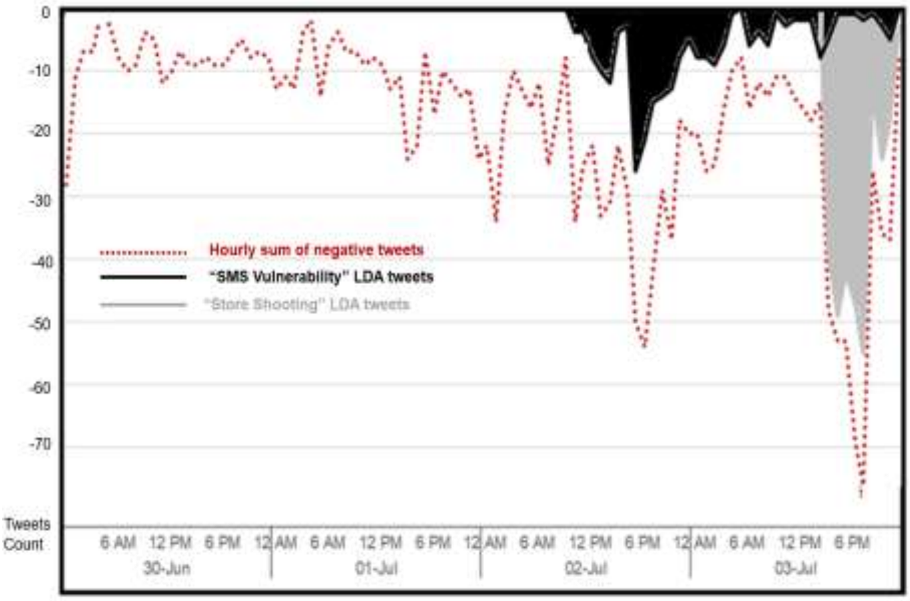


Figure 27: Negative Sentiment Level vs LDA Topics

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 (McCallum, 2002) 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. 27 shows remarkably similar trends 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.

3.11 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, 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. 28 demonstrates the Foreground-Background split.

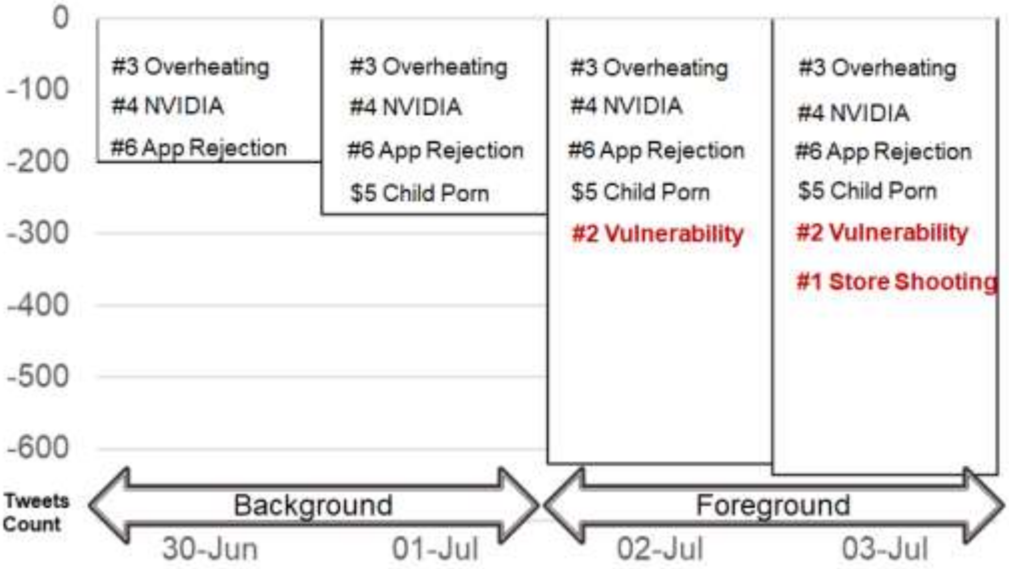


Figure 28: FB-LDA Foreground and Background Periods

As there is no clear guideline for setting the number of topics for the FB-LDA Model, we tried both settings of 2 topics and 5 topics. Fig. 29 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.

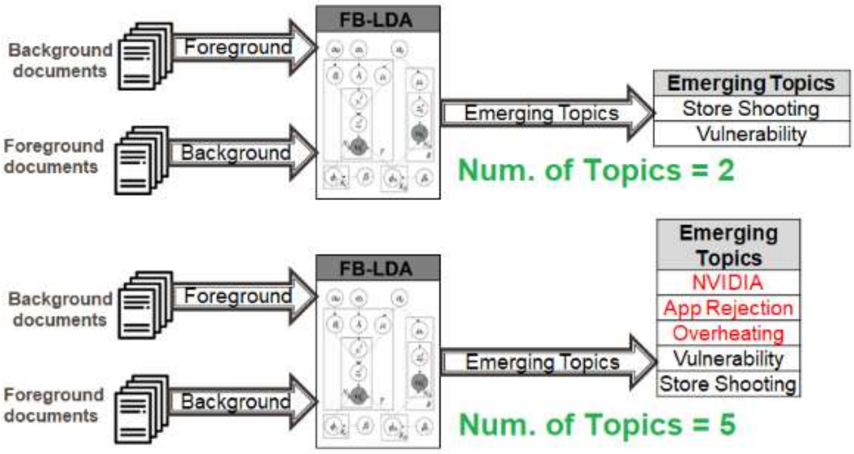


Figure 29: Impact of Number of Topics' Setting on FB-LDA

For instance, the word "Overheat" was listed among the Topic Keywords of the "Store Shooting" topic. RCB-LDA Model was also introduced to ranks the reason candidates. It 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.

The developers of FB-LDA 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, 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.

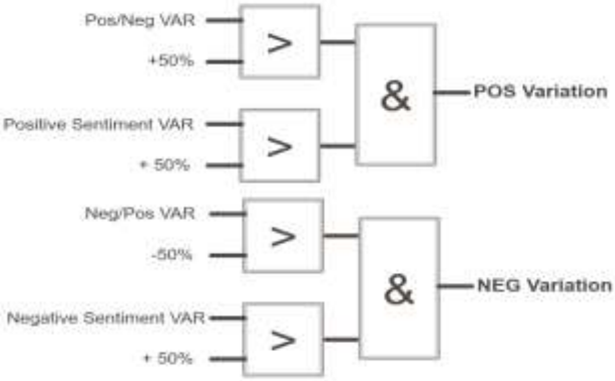


Figure 30: Proposed Sentiment Variations' Detection Method

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. 30.

3.12 Summary

This chapter examined the base methods of Sentiment Variations' Reasoning which we presented in Chapter 2.

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 specific 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 for 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's 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.

Chapter 4 : Frameworks for Emerging Topic Detection*

4.1 Introduction

Finding the right hot scientific topic is a common challenging task for many students and researchers. To illustrate, a PhD student should read a large number of scientific papers on a specific field to identify candidate emerging topics before proposing a dissertation. This time-consuming exercise usually covers multiple years of publications to spot evolution of new topics. During the last two decades, multiple techniques were introduced to handle similar tasks, wherein large sets of timestamped documents are processed by a text mining program to automatically detect emerging topics. In this chapter, we focus only on those techniques that employ Topic Models, which use statistical algorithms to detect topics that appear in texts. Topic Models treat each part of data as a word document. A collection of word documents forms a corpus. Topics can usually be predicted based on some similar words that appear inside each document. Therefore, each document may consist of multiple topics, whereas its dominant topic is the one which is discussed more inside that document.

Some Topic Models are non-probabilistic, like the Latent Semantic Analysis (LSA) (Deerwester et al., 1990) and the Non-Negative Matrix Factorization (NMF) (Lee & Seung, 1999), whereas other Topic Models are probabilistic, like the Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) and the Latent Dirichlet Allocation (LDA) (Blei et al., 2003).

LDA is one of the most used Topic Models because of its good performance and ability to produce coherent outputs for many applications (Alattar & Shaalan, 2021a). LDA represents each document by a distribution of fixed number of topics. Each one of these topics is represented by a distribution of words. Fig. 31 shows a graphical LDA model, which includes three levels of representations.

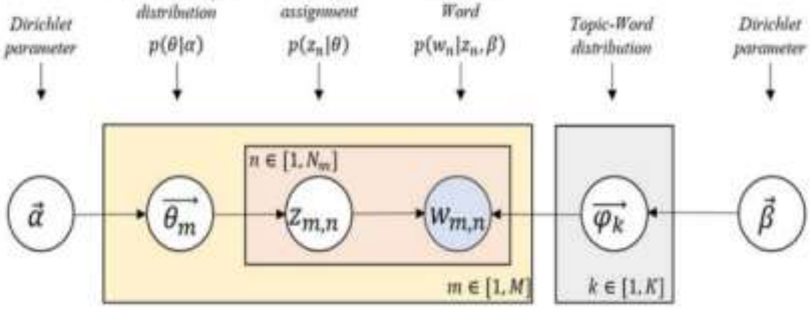


Figure 31: Latent Dirichlet Allocation (LDA) Model (Lee et al., 2018)

* This chapter is derived from the article "Alattar, F. & Shaalan, K., 2021. Emerging Research Topic Detection using Filtered-LDA. MDPI Artificial Intelligence Journal, Volume XX, pp. XX-XX.".

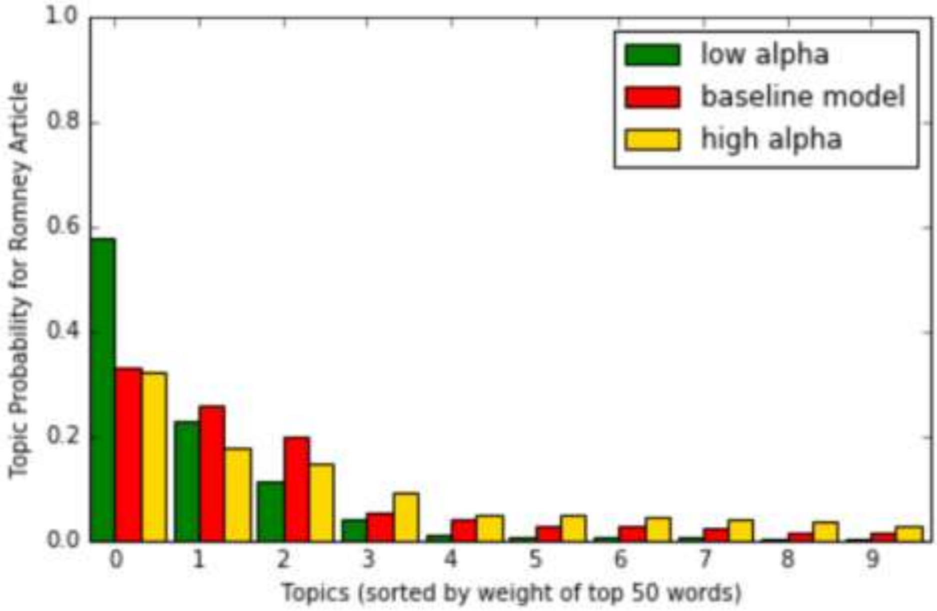


Figure 32: Smoothing impact of LDA hyperparameter Alpha (Hansen, 2016)

Corpus-level representation uses hyperparameters α and β, which are sampled once when a corpus is generated. The document-level representation's variables are $\overrightarrow{\theta m}$, which are sampled once for each document. The word-level representation's variables are $Z_{m,n}$ and W_{mn}, which are sampled once for each word. Given that m is the number of documents in a corpus, and n is the number of words in a document, then D_m is the document number m in the corpus, whereas $\overrightarrow{\theta m}\, is$ the multinomial distribution which includes D_m in the latent topic $Z_{m,n.}$

The hyper-parameter α for this distribution follows the Dirichlet distribution. "K" represents the number of topics which can be either statistically calculated or selected manually be the user. $\overrightarrow{\phi m}$ is the words' multinomial distribution for K topics. This distribution has the hyper-parameter β which follows the Dirichlet distribution. Probability of the word $W_{m,n}$ is decided by p($w_{m,n}$|$z_{m,n}$, β).

Tuning LDA hyperparameters is important for obtaining accurate results. In general, Alpha (α) decides mixture of topics inside a document, whereas Beta (β) decides mixture of words for each topic. For instance, increasing Alpha would increase mixture of topics (Hansen, 2016). Fig. 32 shows example of Wikipedia article on Mitt Romney, which demonstrates the smoothing impact of hyperparameter Alpha. When Alpha is low, wight is assigned to one topic, whereas weight gets distributed among topics when Alpha is high.

Regular Topic Models like LDA use unsupervised learning techniques that are designed to discover discussed topics within text, however they do not have mechanisms to distinguish between old and new topics. As a result, some approaches were developed by various researchers to address the Emerging Topic Detection problem.

4.2 Topic Modeling Base Methods

Given that our research work focuses only on identifying Emerging Topics when two documents are compared, we shall not address here other Emerging Topic Detection problems, like Event Detection for identifying burst topics. In our research problem, an Emerging Topic may exist inside few documents only, and it may not cause a surge that can be discovered through Event Detection techniques. Furthermore, we shall not address the Citation-Based Emerging Topic methods which create clusters from datasets then analyze topics that appear in each cluster. Manual labeling of citation information for each document would require additional human efforts, and our research focuses on detecting all Emerging Topics at an early stage even before they receive high number of citations.

In the following subsections, we briefly address three main Topic Models that are still being used to handle Emerging Topic Detection. These models are the Dynamic Topic Model (DTM), the Partially Labeled Dirichlet Allocation (PLDA), and the Foreground-Background Latent Dirichlet Allocation (FB-LDA). Later, we test these models on a real-life dataset to examine their performance, then we introduce our proposed Filtered-LDA frameworks to overcome limitations of existing models.

4.2.1 Dynamic Topic Model

DTM was developed to examine evolution of topics inside a large set of documents. It is a probabilistic time series model that utilizes multinomial distribution to represent topics. Unlike static models, posterior inference for a dynamic model is intractable, therefore DTM utilizes wavelet regression and Kalman filters to approximate inference (Blei & Lafferty, 2006).

Blei & Lafferty (2006) applied DTM on a large set of Science Journal documents from year 1880 to 2000. They presented the results of both wavelet regression and Kalman filters separately. Both approximation methods could track the variation of target topic frequencies over time, with differences related to smoothing some topic curves and spikes due to dissimilarities in these approximation methods.

In practice, a DTM program divides the complete time span into multiple time points, and the specified number of documents are detected for each time point. During training process, the program keeps adding a new document to analyze, and finally, the output produces distribution of each topic over the series of time points. Fig. 33 shows a DTM as used by tomotopy Topic Modeling toolkit, where documents are split into two time slots: first time slot includes Background documents which represent old documents, whereas second time slot includes Foreground documents which represent newer documents. Documents can be divided into multiple number of time slots.

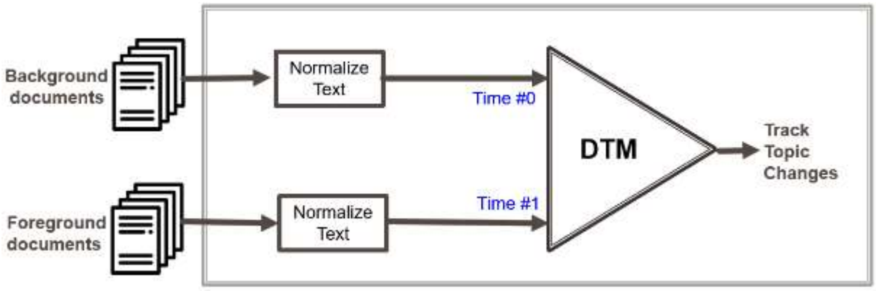


Figure 33: Dynamic Topic Model (DTM)

Given that DTM limits its topic discovery to the prespecified number of topics, it keeps tracking these topics until the end of the time series. Therefore, in case a new topic emerges after reaching this prespecified number, it is expected that DTM will not be able to detect this new topic. Increasing DTM number of topics does not help its ability to detect emerging topics because the algorithm tries to detect same number of topics from each timeslot. We shall demonstrate this limitation later in Section 4.4 through an experiment.

4.2.2 Partially Labeled Dirichlet Allocation

PLDA was introduced to enhance human interpretability of Topic Models' outputs. PLDA is a semi-supervised model, which utilizes available manual topic labels for part of a set of documents. The model scans documents and links discovered topics with their associated label. It also identifies topics that are not represented by any label. Ramage et al. (2011) applied PLDA on a large set of PhD dissertations dataset, and it could group discovered topics according to the prespecified labels. Moreover, it could also identify those topics which are not represented by any label.

Fig. 34 shows a PLDA Model as used by tomotopy Topic Modeling Toolkit, where labels are provided for the Background documents, which represent old documents, then the model would be trained on the unlabeled Foreground documents, which represent newer documents' set.

By using PLDA for solving the Emerging Topic Detection problem, new topics should be detected once the user knows the labels of old topics. This technique looks promising, however when you analyze its mechanism, you realize that discovered unlabeled topics are mainly the global topics which exist in most of the documents.

As a result, it is not expected from PLDA to efficiently detect Emerging Topics when these are only discussed locally in some documents. This expectation shall be validated later in Section 4.4 through an experiment using a labeled dataset.

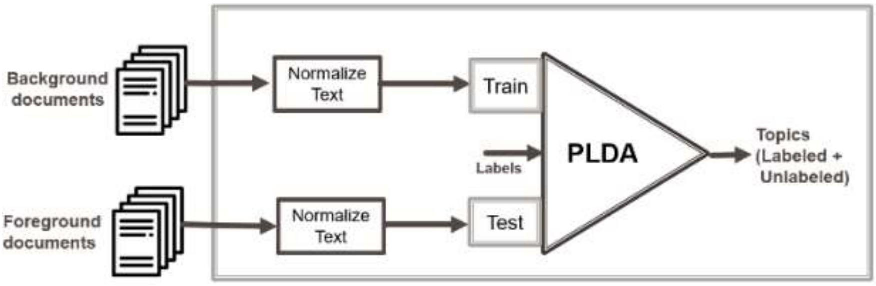


Figure 34: PLDA for Background and Foreground document

4.2.3 FB-LDA Topic Model

FB-LDA Model aims to discover emerging topics when two sets of documents are compared. The first set of documents is called the Background, whereas the second set is called the Foreground. The Foreground documents appear during the Foreground period in which we are interested in identifying Emerging Topics that did not appear in the past. The Background period is the period which ends just before the start of the Foreground period, and its duration is double of the Foreground period. FB-LDA Model detects new topics from the Foreground after removing all topics that are discussed in the Background.

FB-LDA utilizes regular LDA Model with Gibbs Sampling to detect topics from both Background and Foreground documents. However, after identifying the topics of the Background documents, it starts classifying the Foreground topics based on their similarity with the Background topics. If a Foreground topic is not similar to a Background topic, it is identified as an Emerging Topic. Fig. 35 shows a representation of FB-LDA Topic Model. This technique can also be applied to timestamped document sets by slicing the time span into either equal or unequal time slots, then detecting the new topics from any given Foreground timeslot when compared to Background time slots.

Tan et al. (2014) applied FB-LDA on a set of timestamped scientific paper abstracts by splitting them into two groups. The first group represents the recent abstracts which were written during the last 3 years in the dataset, whereas the second group includes the rest of abstracts which were written during the earlier 10 years. Then FB-LDA was used to discover the hot scientific topics which are discussed in the Foreground set, but not discussed in the Background set.

In this chapter, we shall repeat the same experiment of the developers of FB-LDA to analyze its outcomes. We shall also apply FB-LDA on an additional dataset to verify our findings.

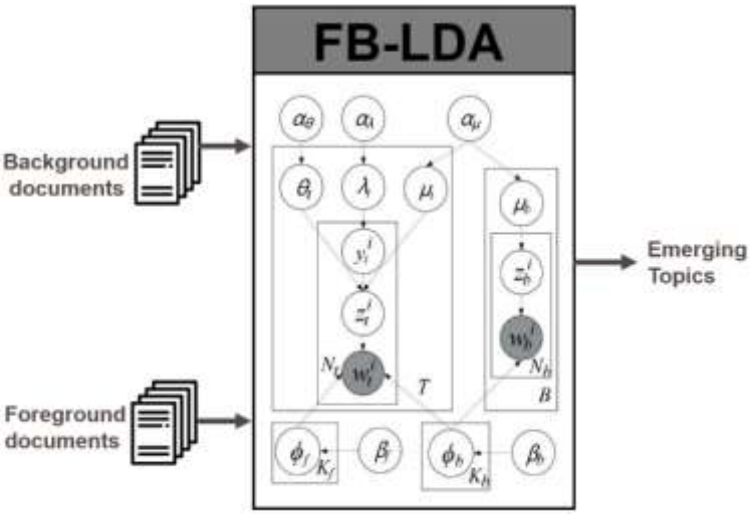
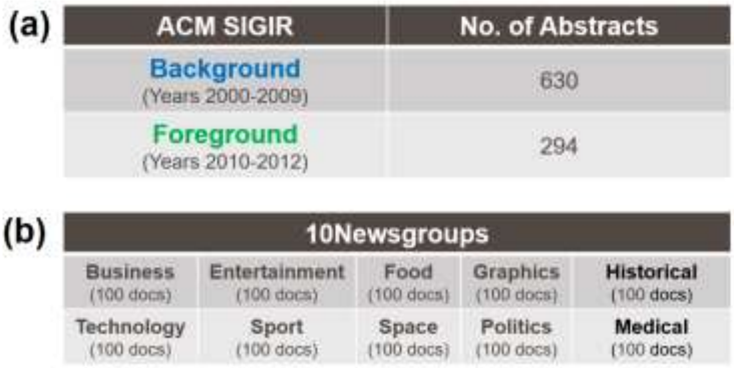


Figure 35: FB-LDA Model

4.3 Datasets

For our experiments, we shall use two different medium-size datasets. The first one is the ACM SIGIR dataset (Tan, 2018) which we selected so that we can compare our results with Tan et al. (2014) who selected the same dataset to test their FB-LDA model. It consists of 924 scientific paper abstracts about Information Retrieval separated into two groups as illustrated in Table 4. First group represents Background period from year 2000 to 2009, whereas second group includes Foreground documents from year 2010 to 2012.



(a)

ACM SIGIR	No. of Abstracts
Background (Years 2000-2009)	630
Foreground (Years 2010-2012)	294

(b)

10Newsgroups				
Business (100 docs)	Entertainment (100 docs)	Food (100 docs)	Graphics (100 docs)	Historical (100 docs)
Technology (100 docs)	Sport (100 docs)	Space (100 docs)	Politics (100 docs)	Medical (100 docs)

Table 4: Datasets (a) SIGIR and (b) 10Newsgroups

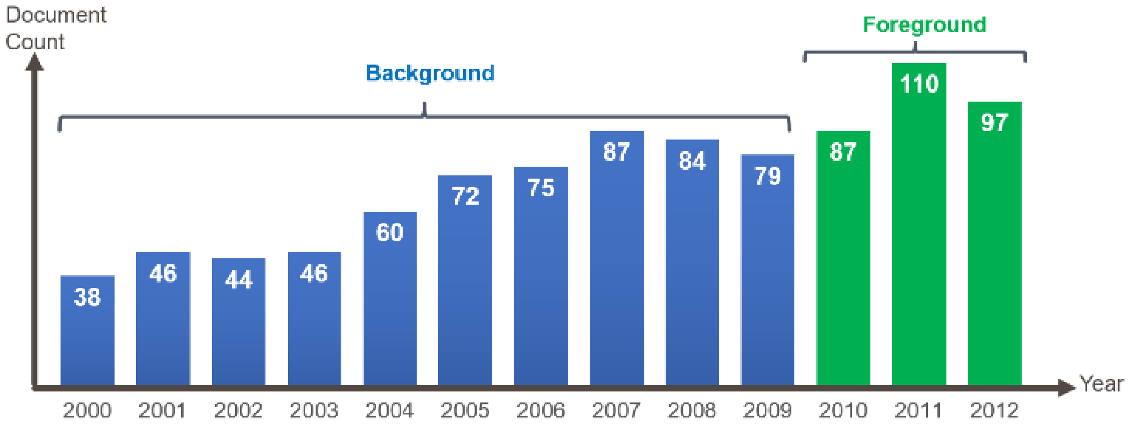


Figure 36: Distribution of ACM SIGIR 2000-2012 documents

We manually labeled the timestamp of each abstract by recording its publication year using a simple Google Scholar search for each paper title. Fig. 36 shows distribution of abstracts from year 2000 to 2012, where Background documents are from year 2000 to 2009, whereas Foreground documents are from year 2010 to 2012.

To categorize the topics which are discussed in SIGIR dataset, we read all the 924 abstracts, and we manually labeled the dominant topic of each abstract. As explained earlier, each document consists of multiple topics, and each topic has its own probability inside that document. The dominant topic for each document is that topic which has the highest probability. We ignored topics that do not appear in more than 3 abstracts because these do not represent a scientific trend. Finally, we grouped all SIGIR abstracts based on their common labels as follows:

Background topics which also appear in Foreground:

- User Behavior
- Question Answering
- Search Engines Queries
- Relevance Feedback
- Recommendation Items
- Average Precision Evaluation Metrics
- Learning to Rank
- Image Retrieval
- Web Search Pages

Background topics which do not appear in Foreground:

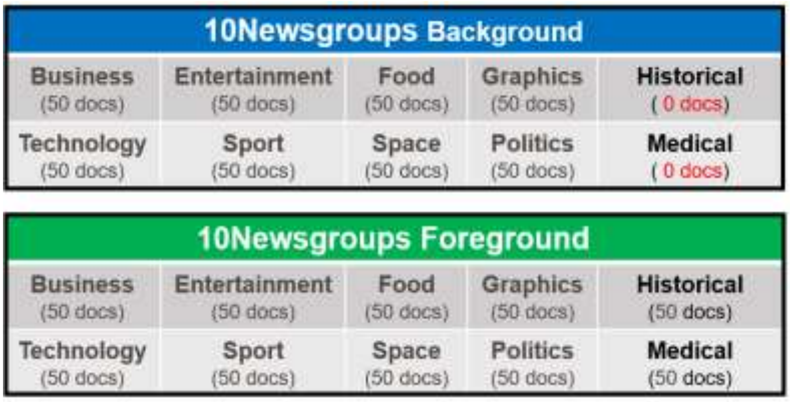
- Topic Detection and Tracking Event
- Document Clustering
- Language Translation Models

Emerging topics which mainly appear in Foreground:

- Search Result Diversification
- Feature Space Hashing
- Social Network Twitter
- Search Result Cache Invalidation
- Temporal Indexing

Given that– in practice – an Emerging Topic could appear in one or two papers during Background period, we kept such topics in the Emerging Topics list if they appear at least in 3 documents during the Foreground period. For example, we consider Temporal Indexing as an Emerging Topic because it appears in 3 abstracts during Foreground period, although it also appears once in the Background period.

The second dataset is the 10Newsgroups (Lang, 2008) which includes 1000 classified news messages on 10 different topics as shown in Table 4. The second dataset is used for examining the ability of base methods to detect emerging topics by manually splitting the 10Newsgroups documents into two sets. First set represents Background documents, and second set represents Foreground documents. We randomly selected documents from two topics that represent Emerging Topics and inserted these documents in the Foreground group only.



10Newsgroups Background				
Business (50 docs)	Entertainment (50 docs)	Food (50 docs)	Graphics (50 docs)	Historical (0 docs)
Technology (50 docs)	Sport (50 docs)	Space (50 docs)	Politics (50 docs)	Medical (0 docs)

10Newsgroups Foreground				
Business (50 docs)	Entertainment (50 docs)	Food (50 docs)	Graphics (50 docs)	Historical (50 docs)
Technology (50 docs)	Sport (50 docs)	Space (50 docs)	Politics (50 docs)	Medical (50 docs)

Table 5: Background & Foreground of 10Newsgroups dataset

Multiple combinations of Emerging Topics are used to verify our findings. For example, Table 5 shows the split of Background and Foreground documents, wherein we included Historical and Medical documents only in the Foreground.

4.4 Experiment Using Base Methods

We used the datasets of Section 4.3 to perform Python experiments for analyzing outputs of the three base methods: DTM, PLDA, and FB-LDA.

4.4.1 Experimental Setup

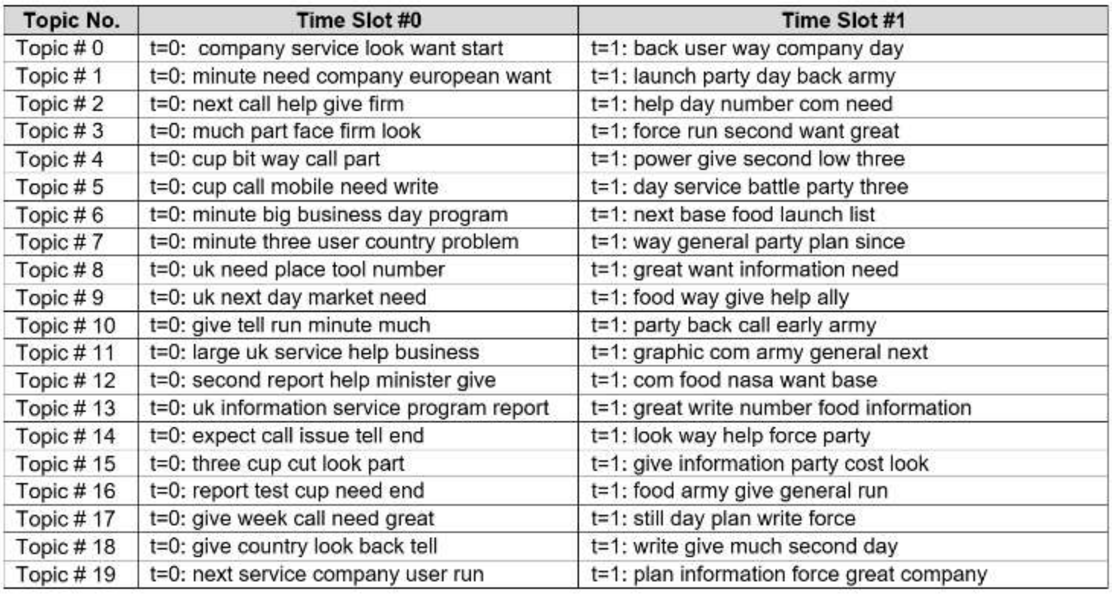
We used a Dell Inspiron 7370 laptop with Intel$^{(R)}$ Core$^{(TM)}$ i7-8550U CPU @ 1.99 GHz Processor, Installed RAM is 16.0 GB, Windows 10 Home version 20H2 64-bit operating system. Python version 3.8.3 is used with Jupyter Notebook server version 6.1.4.

We applied DTM on SIGIR dataset by splitting it into two timepoints: Time #0, which includes all scientific abstracts from year 2000 to 2009, and Time #1, which includes all abstracts from year 2010

Before applying the Topic Modeling algorithms, we used Genism library to carry out simple text normalization and data preprocessing steps which include (1) stripping html tags, (2) removing accent characters, (3) expanding contractions, (4) removing special characters, (5) removing digits, (6) removing extra new lines, (7) removing extra white space, (8) removing stop words, (9) lowercasing each word (10) tokenization, wherein sentences and words are extracted from each document, (11) lemmatizing words by converting all verbs to present tense and grouping different forms of a noun to a common form, and (12) stemming which reduces each word to its root.

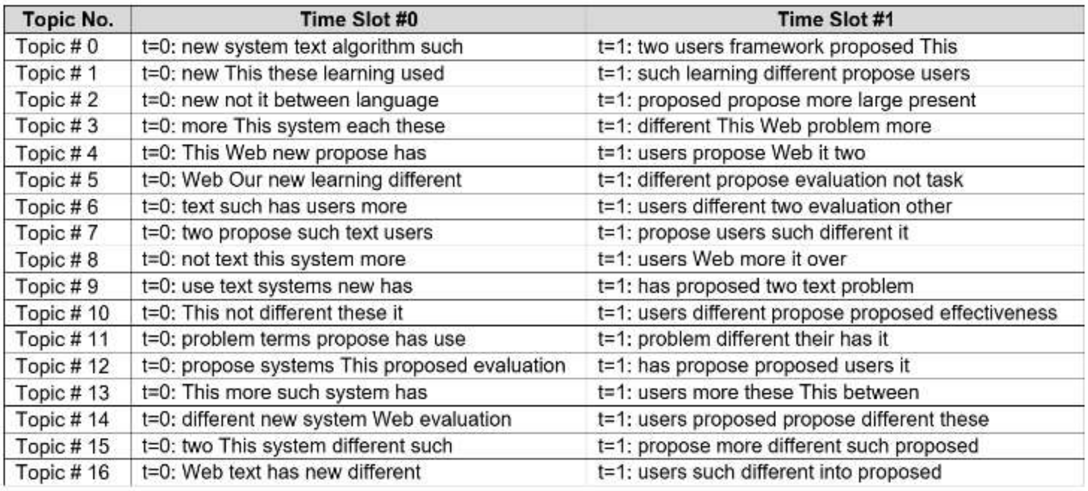
To apply DTM and PLDA, we used Python wrapper for tomotopy, which is a Topic Modeling Tool written in C++ based on Gibbs-sampling. For FB-LDA, we used the Python Gibbs-sampling implementation program (Tan, 2018), which is provided by the developers of FB-LDA.

4.4.2 Dynamic Topic Model



Topic No.	Time Slot #0	Time Slot #1
Topic # 0	t=0: company service look want start	t=1: back user way company day
Topic # 1	t=0: minute need company european want	t=1: launch party day back army
Topic # 2	t=0: next call help give firm	t=1: help day number com need
Topic # 3	t=0: much part face firm look	t=1: force run second want great
Topic # 4	t=0: cup bit way call part	t=1: power give second low three
Topic # 5	t=0: cup call mobile need write	t=1: day service battle party three
Topic # 6	t=0: minute big business day program	t=1: next base food launch list
Topic # 7	t=0: minute three user country problem	t=1: way general party plan since
Topic # 8	t=0: uk need place tool number	t=1: great want information need
Topic # 9	t=0: uk next day market need	t=1: food way give help ally
Topic # 10	t=0: give tell run minute much	t=1: party back call early army
Topic # 11	t=0: large uk service help business	t=1: graphic com army general next
Topic # 12	t=0: second report help minister give	t=1: com food nasa want base
Topic # 13	t=0: uk information service program report	t=1: great write number food information
Topic # 14	t=0: expect call issue tell end	t=1: look way help force party
Topic # 15	t=0: three cup cut look part	t=1: give information party cost look
Topic # 16	t=0: report test cup need end	t=1: food army give general run
Topic # 17	t=0: give week call need great	t=1: still day plan write force
Topic # 18	t=0: give country look back tell	t=1: write give much second day
Topic # 19	t=0: next service company user run	t=1: plan information force great company

Table 6: DTM Results for 10Newsgroups dataset



Topic No.	Time Slot #0	Time Slot #1
Topic # 0	t=0: new system text algorithm such	t=1: two users framework proposed This
Topic # 1	t=0: new This these learning used	t=1: such learning different propose users
Topic # 2	t=0: new not it between language	t=1: proposed propose more large present
Topic # 3	t=0: more This system each these	t=1: different This Web problem more
Topic # 4	t=0: This Web new propose has	t=1: users propose Web it two
Topic # 5	t=0: Web Our new learning different	t=1: different propose evaluation not task
Topic # 6	t=0: text such has users more	t=1: users different two evaluation other
Topic # 7	t=0: two propose such text users	t=1: propose users such different it
Topic # 8	t=0: not text this system more	t=1: users Web more it over
Topic # 9	t=0: use text systems new has	t=1: has proposed two text problem
Topic # 10	t=0: This not different these it	t=1: users different propose proposed effectiveness
Topic # 11	t=0: problem terms propose has use	t=1: problem different their has it
Topic # 12	t=0: propose systems This proposed evaluation	t=1: has propose proposed users it
Topic # 13	t=0: This more such system has	t=1: users more these This between
Topic # 14	t=0: different new system Web evaluation	t=1: users proposed propose different these
Topic # 15	t=0: two This system different such	t=1: propose more different such proposed
Topic # 16	t=0: Web text has new different	t=1: users such different into proposed

Table 7: DTM Results for SIGIR 2000-2012 dataset

We applied DTM on SIGIR dataset by splitting it into two timepoints: Time #0, which includes all scientific abstracts from year 2000 to 2009, and Time #1, which includes all abstracts from year 2010 to 2012.

The number (K=17) is selected as we are aware that the number of main topics in the dataset is 17 based on our manual labeling process. Our program ignored topics which do not appear in more than 3 documents. However, as shown in Table 6, DTM failed to detect any Emerging topic. A similar result was obtained when number of topics was doubled to become (K=34).

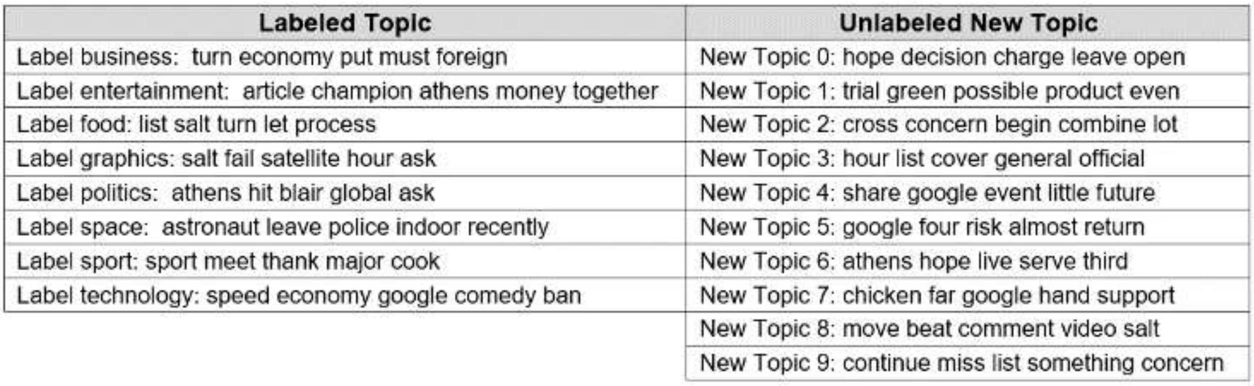
We also applied DTM on the 10Newsgroups dataset by splitting it into two timepoints: Time #0, which includes all Background documents of Table 5, and Time #1, which includes all Foreground documents of same table. We selected (K=10) as we are aware that the number of main topics in the dataset is 10. However, as shown in Table 7, DTM also failed to detect any Emerging topic. A similar result was obtained when we doubled number of topics to become (K=20). These two experiments on DTM confirm our expectations that DTM cannot detect Emerging Topics when these do not appear in the background time slot.

4.4.3 Partially Labeled Dirichlet Allocation

The methodology of this experiment is to provide the labels of background dataset, hoping that PLDA will be able to detect unlabeled topics in the Foreground dataset. We provided the labels of Background documents to PLDA, then we applied it on the Foreground dataset, which includes Emerging Topics in addition to Background topics.

For the 10Newsgroups dataset, we selected (K=10) for the number of PLDA topics as we are aware that the number of main topics in the dataset is 10. However, as shown in Table 8, PLDA failed to detect any Emerging topic as none of the new unlabeled topics is related to Historical or Medical topics.

Although PLDA is good for tracking all topics which are detected during the training process, its detection of unlabeled topics failed to discover Emerging Topics as it could only detect new sub-topics that are related to labeled topics.

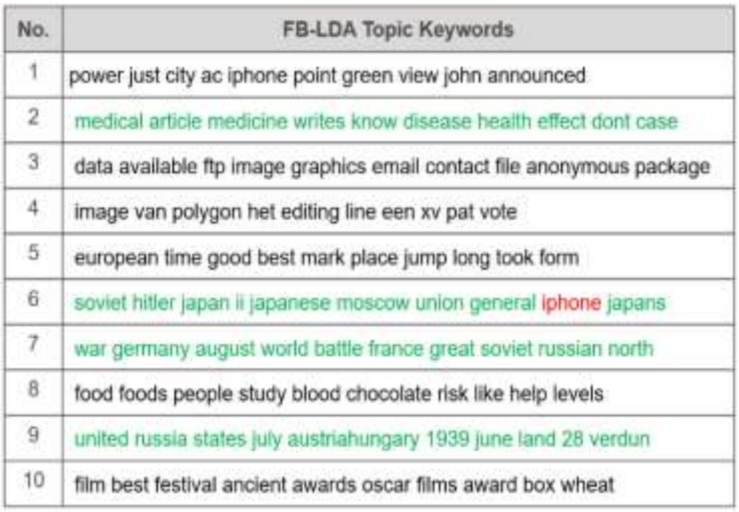


Labeled Topic	Unlabeled New Topic
Label business: turn economy put must foreign	New Topic 0: hope decision charge leave open
Label entertainment: article champion athens money together	New Topic 1: trial green possible product even
Label food: list salt turn let process	New Topic 2: cross concern begin combine lot
Label graphics: salt fail satellite hour ask	New Topic 3: hour list cover general official
Label politics: athens hit blair global ask	New Topic 4: share google event little future
Label space: astronaut leave police indoor recently	New Topic 5: google four risk almost return
Label sport: sport meet thank major cook	New Topic 6: athens hope live serve third
Label technology: speed economy google comedy ban	New Topic 7: chicken far google hand support
	New Topic 8: move beat comment video salt
	New Topic 9: continue miss list something concern

Table 8: PLDA Results for 10NewsGroups dataset

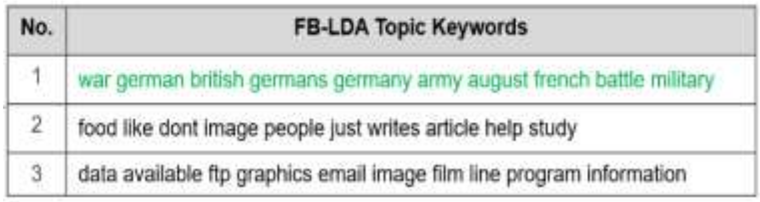
4.4.4 FB-LDA Topic Model

We applied FB-LDA on the 10Newsgroups dataset and selected (K=10) as the number of topics. As clear from Table 9, the model could detect Emerging Topics successfully, however the keywords of 60% of detected topics do not belong to emerging Historical and Medical topics.



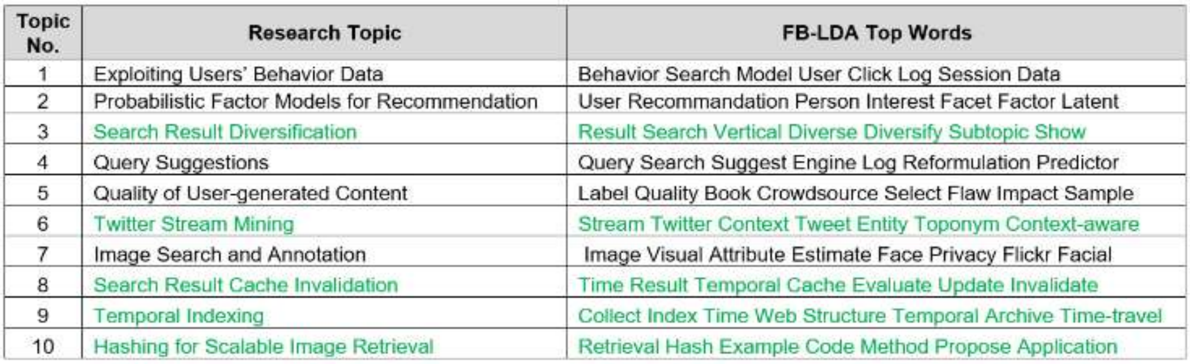
No.	FB-LDA Topic Keywords
1	power just city ac iphone point green view john announced
2	medical article medicine writes know disease health effect dont case
3	data available ftp image graphics email contact file anonymous package
4	image van polygon het editing line een xv pat vote
5	european time good best mark place jump long took form
6	soviet hitler japan ii japanese moscow union general iphone japans
7	war germany august world battle france great soviet russian north
8	food foods people study blood chocolate risk like help levels
9	united russia states july austriahungary 1939 june land 28 verdun
10	film best festival ancient awards oscar films award box wheat

Table 9: FB-LDA (K=10) Results for 10NewsGroups dataset (correct Emerging Topics are marked in green)



No.	FB-LDA Topic Keywords
1	war german british germans germany army august french battle military
2	food like dont image people just writes article help study
3	data available ftp graphics email image film line program information

Table 10: FB-LDA (K=3) Results for 10NewsGroups dataset (correct Emerging Topics are marked in green)



Topic No.	Research Topic	FB-LDA Top Words
1	Exploiting Users' Behavior Data	Behavior Search Model User Click Log Session Data
2	Probabilistic Factor Models for Recommendation	User Recommandation 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

Table 11: FB-LDA results for SIGIR dataset derived from (Tan, et al., 2014). We marked the correctly detected Emerging Topics in green color.

For example, the keywords of the first topic belong to the Technology topic, whereas the keywords of the third topic belong to the Graphics topic. Furthermore, we noticed that the word "phone" from the background Technology topic appeared in the emerging Historical topic. When number of topics is reduced to 3, the model could detect only one Emerging Topic as shown in Table 10.

In our experiment, FB-LDA is also applied on SIGIR dataset by selecting (k=10) as the number of topics, and we could obtain similar results to the ones published by the developers of FB-LDA as shown in Table 11. However, it is noticed that 50% of detected topics are not Emerging Topics. As a result, we conclude that FB-LDA can detect some of the emerging topics when number of topics is selected correctly, though the developers of FB-LDA did not provide clear guidelines for selecting the setpoint of number of topics. However, the output of FB-LDA showed some Background topics most of the time, and in few cases, it also showed a Background word along the top keywords of an Emerging Topic.

4.5 Filtered-LDA Model

Our experiments of Section 4.4 revealed that both DTM and PLDA could not detect emerging topics from our two datasets efficiently, whereas FB-LDA could do the job with some limitations. In this section, we introduce our novel model to overcome limitations of base methods. We called the new model "Filtered-LDA" as it directs its Topic Modeling components to filter out old topics and keep new topics only.

Fig. 37 presents the proposed Filtered-LDA Model, which attempts to emulate human approach for discovering new topics from a large set of documents. A human would first skim through all documents to isolate the ones which he/she thinks they contain new topics. Then he/she would only go through the isolated set of documents to identify high-frequency emerging topics.

With this simple technique, the impact of isolation/clustering errors would be insignificant because the wrongly classified documents would not be among the high-frequency emerging topics.

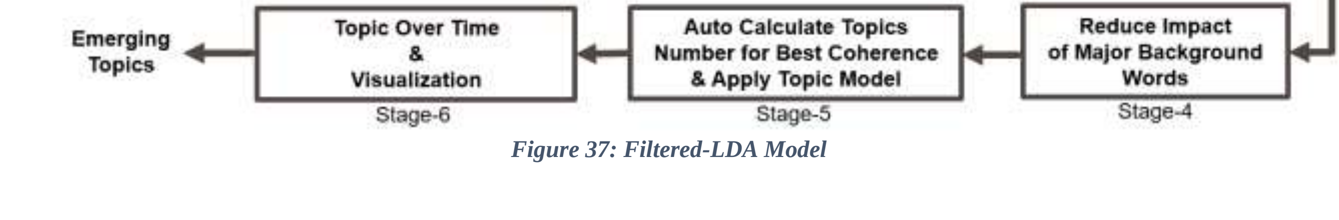

Full Dataset
(Background & Foreground)
Stage-1
Normalize Text & Filter Out Foreground No-Emerging-Words documents
Stage-2
Auto Calculate Topics Number for Best Perplexity & Apply Topic Model
Stage-3
Filter docs by Keeping only Foreground docs with Emerging Topics
Reduce Impact of Major Background Words
Stage-4
Auto Calculate Topics Number for Best Coherence & Apply Topic Model
Stage-5
Topic Over Time & Visualization
Stage-6
Emerging Topics

Figure 37: Filtered-LDA Model

Nevertheless, this novel model introduces additional measures to ensure high performance of Emerging Topic Detection by genuinely reducing the chance of detecting old topics.

First stage of the model aims to reduce number of Background topics from the Foreground documents to minimum by ignoring all Foreground documents that do not contain new words when compared to Background documents. In practice, Emerging Topics could appear once or twice in the Background document set, but they do not appear as a trend in the Background duration. The threshold of the number of documents that represent a trend is a variable that shall be defined by the user. For instance, if a user is looking for a completely new topic that never appeared in the Background, then the threshold setpoint is 0. However, if a user accepts Emerging Topics even if they appear once or twice in the Background then the threshold setpoint is 2. In our experiments, we selected multiple options of hot trend threshold between 0 to 10, and we could achieve consistent results with accuracy above 90%.

The first stage of the model also ignores words that have low frequency in the Background documents. This aims to reduce possibility of mixing Foreground Emerging Topics with Background Topics due to presence of some low frequency words in the Background dataset. Second stage uses Topic Modeling for clustering topics in the complete dataset. To select optimum number of topics for the Topic Model, usually number of topics that gives highest Coherence Score is selected. As explained by Roder, Both, & Hinneburg (2015), there are multiple methods for measuring the topic's Coherence Score, and the C_V measurement showed the best performance. Therefore, in our experiments we select the C_V measurement which uses a sliding window, normalized Pointwise

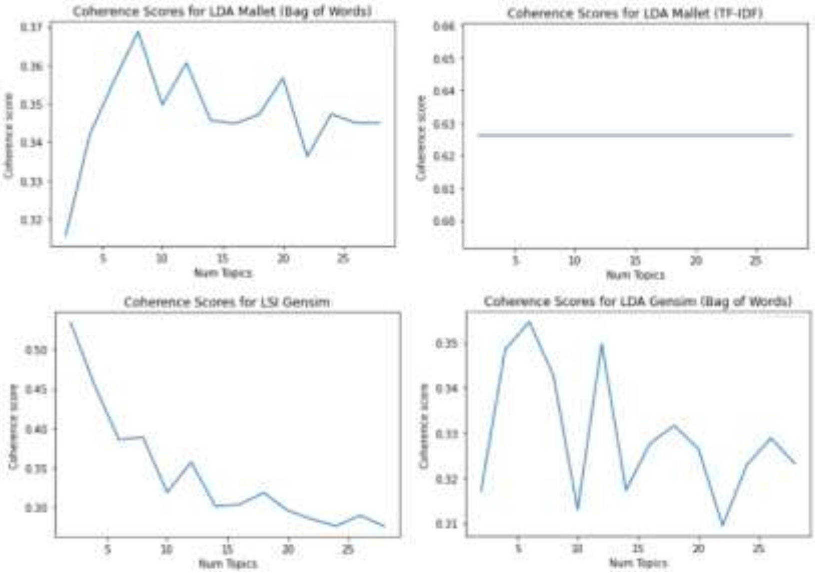


Figure 38: Topic Models' Coherence Scores curves for SIGIR Dataset

Mutual Information (NPMI), and the cosine similarity. However, we noticed that highest Coherence Score does not guarantee best representation of dataset. Fig. 38 shows an example in which Topic Models are used to analyze our SIGIR dataset. We know from our manual review of the dataset that the minimum number of topics is 17 as explained earlier in Section 4.3. However, regular Topic Models that use Bag of Words (BoW) text representation showed highest Coherence Scores for topic numbers that are much lower than 17. For instance, Gensim LDA showed highest Coherence Score when (K=8), which means that remaining 9 topics would not be captured. Hence, at this stage, we shall select the number of topics "K" that makes our model more fit for the test dataset or the held-out data without caring much about the human interpretability of the model because this output is not the one which will be finally presented to the user. Therefore, this stage focuses on the Perplexity Score of the Topic Model, rather than focusing on the Coherence score. There are multiple ways for carrying out Stage-2 task efficiently. We shall propose two different options of frameworks that can ensure good clustering performance for our model.

Third stage identifies all topics that do not have documents in the Background dataset and topics that form a trend in Foreground dataset but do not form a trend in Background dataset. As explained earlier, the number of documents that represents a trend is a variable that the user shall select in advance. For our experiment, a trend or a hot topic shall be discussed at least in 3 documents. As a result, the output of stage 3 is all Foreground documents discussing topics that do not represent major Background topics. To ensure that keywords of major Background topics do not appear in the Emerging Topics keywords, the fourth stage filters out high frequency Background words from the obtained Foreground documents of third stage.

Fifth stage automatically selects optimum number of topics "K" for the filtered Foreground documents by calculating the Coherence Scores for a wide range of "K" values so that the output of the Topic Model at the sixth stage can be easily interpreted by the user.

Final stage presents detected Emerging Topics in multiple forms including drawing Topic Over Time curves, topic keywords list, topic keywords' Wordcloud, visual representation of topic probabilities, visual representation of topic similarities and overlap, and finding the most representative document for each topic. Automatic topic labeling (Mei et al., 2007) is also used at this stage to select best few keywords that represent the topic. This ensures better interpretation of discovered topics and that the user can monitor evolution of topics over time.

4.5.1 First Framework

Low Perplexity score for a model ensures that it represents the held-out data well (Yarnguy & Kanarkard, 2018), hence “K” that gives a low LDA Perplexity score shall be selected for Stage 2 of the model. However, to reduce the computational power requirements for this step, an alternative method is chosen by using the Hierarchical Dirichlet Processes (HDP) as shown in Fig. 39 which presents the first framework. HDP has low Perplexity scores regardless of the number of topics (Teh et al., 2004) when compared to LDA as shown in Fig. 40.

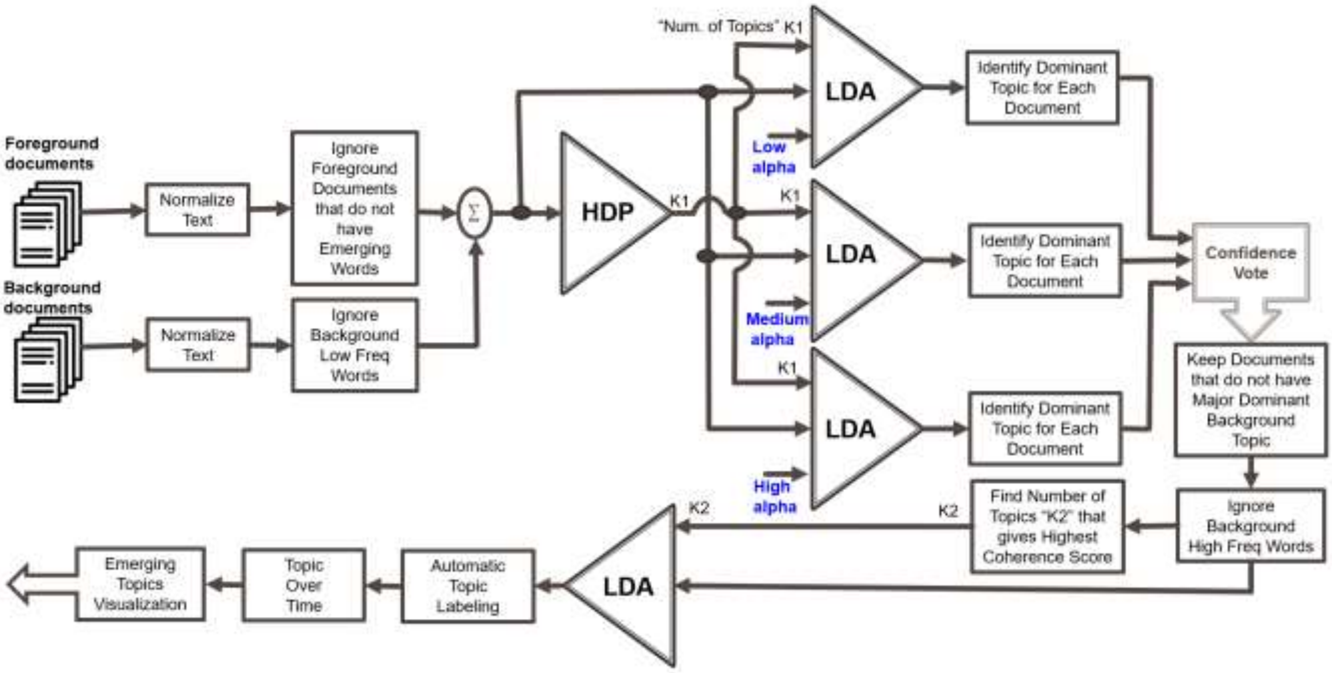

Foreground documents
Background documents
Normalize Text
Normalize Text
Ignore Foreground Documents that do not have Emerging Words
Ignore Background Low Freq Words
Σ
HDP
K1
"Num. of Topics" K1
LDA
Low alpha
K1
LDA
Medium alpha
K1
LDA
High alpha
Identify Dominant Topic for Each Document
Identify Dominant Topic for Each Document
Identify Dominant Topic for Each Document
Confidence Vote
Keep Documents that do not have Major Dominant Background Topic
Ignore Background High Freq Words
Find Number of Topics "K2" that gives Highest Coherence Score
K2
K2
LDA
Automatic Topic Labeling
Topic Over Time
Emerging Topics Visualization

Figure 39: Emerging Topic Detection Filtered-LDA Framework #1

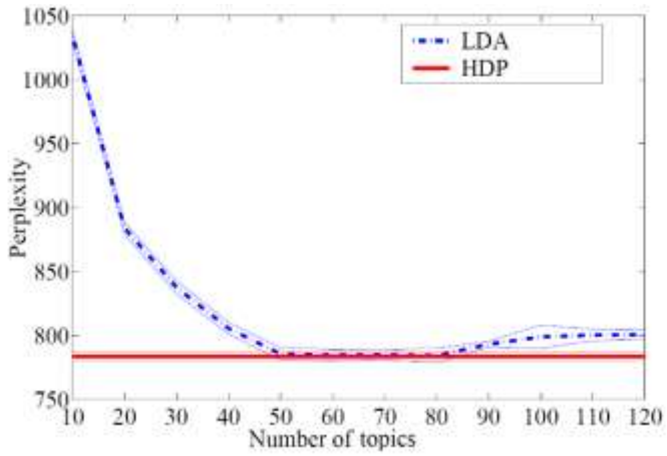


Figure 40: Comparison of LDA and HDP Perplexity (Limsettho et al., 2014)

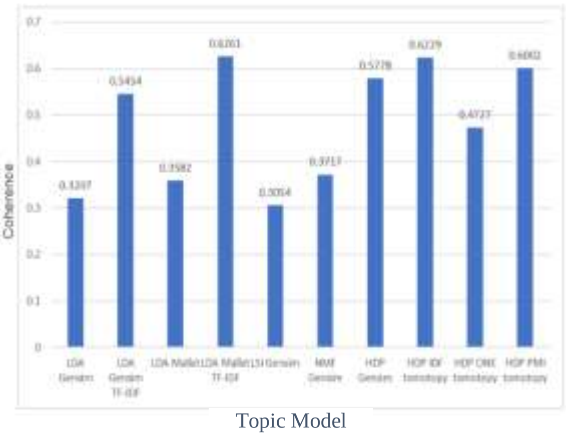


Figure 41: Topic Models’ Coherence Scores for SIGIR Dataset at K=20

Moreover, HDP also provides good Coherence Scores when compared to other models as shown in the example of Fig. 41 in which we compared various types of Topic Models. It is noticed that HDP scores are higher than Models that use BoW text representation.

Employing HDP for the second stage also solves the problem of manual selection of the number of topics "K" as HDP automatically calculates optimum number of topics for the document set. However, the precision of HDP Model depends on the selected Term Weighting Scheme (TM) (Wilson & Chew, 2010), therefore a comparison for these schemes shall be carried out for the dataset, and the TM that provides highest Coherence Score shall be automatically selected.

We selected LDA Model for the actual clustering work because we could achieve slightly higher accuracy using LDA when compared to HDP results. Similar findings were confirmed earlier by the work of Limsettho et al. (2014).

Cascaded LDA models are used to look at the dataset from multiple distances by choosing different value of Alpha hyperparameter for each LDA block so that Emerging Topics would not be merged by mistake with other old topics due to wrong smoothing factor. We used in Fig. 39 three cascaded LDA models, one with low Alpha (α=1), second with medium Alpha (α=50), and third with high Alpha (α=100). These settings are used with Gensim wrapper for Mallet toolkit. Of course, cascaded LDA block can have additional LDA models to cover wider range of Alpha, including very high settings. In case program speed is not important for an application, we recommend using high number of cascaded LDA that covers wide range of Alpha settings to enhance the zooming effect of the model.

A simple Confidence Voting process follows the output of cascaded LDA to select those documents that appeared multiple times. To illustrate, if a document is identified as an Emerging Topic's document at the output of LDA Models of both low Alpha and Medium Alpha, it shall be considered as an Emerging Topic's document.

Afterwards, high frequency Background words are removed from all selected documents to ensure that final Emerging Topics do not use common Background keywords.

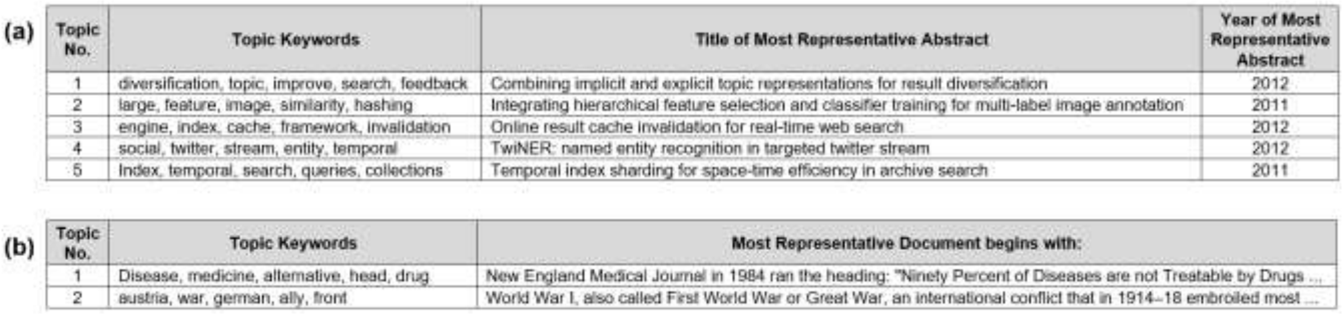
Next stage includes a standard LDA Model with optimum Coherence score as described earlier in Stage 6 of the model.

Finally, Topic Over Time is drawn using LDA calculated probabilities for topics during each time slot. Each document is automatically labeled with a single topic based on its dominant topic which has the highest probability in that document. Then each topic is tracked over time based on the number of documents where it has highest probability.

To implement the framework, Python is used along with multiple packages including Gensim (Rehurek, 2021) for Topic Modeling, and pyLDAvis for topic visualization. Given that current Gensim version supports only Variational Bayes Sampling for LDA, we used Python wrapper for Mallet (McCallum, 2002) toolkit to implement LDA with Gibbs Sampling because it showed better results as illustrated in Fig. 40. We also used Python wrapper for tomotopy toolkit to implement HDP and automatic topic labeling.

To test the framework, experiments are carried out on both SIGIR and 10Newsgroups datasets. For the HDP Model in Stage 2, the Term Weighting Scheme that provides highest coherence score for the model is automatically selected to ensure good precision. For example, in Fig. 41, the Inverse Document Frequency term weighting TM IDF showed highest coherence score for the SIGIR dataset when compared to the Pointwise Mutual Information term weighting TM PMI and the TM ONE which considers every term equal.

The automatic numbers of topics “K” for HDP at Stage 2 for SIGIR and 10Newsgroups datasets were 100 and 92 respectively, whereas the number of LDA topics “K” at Stage 5 were 5 and 2 respectively. Table 12 shows the Emerging Topics of both datasets.



(a)

Topic No.	Topic Keywords	Title of Most Representative Abstract	Year of Most Representative Abstract
1	diversification, topic, improve, search, feedback	Combining implicit and explicit topic representations for result diversification	2012
2	large, feature, image, similarity, hashing	Integrating hierarchical feature selection and classifier training for multi-label image annotation	2011
3	engine, index, cache, framework, invalidation	Online result cache invalidation for real-time web search	2012
4	social, twitter, stream, entity, temporal	TwiNER: named entity recognition in targeted twitter stream	2012
5	Index, temporal, search, queries, collections	Temporal index sharding for space-time efficiency in archive search	2011

(b)

Topic No.	Topic Keywords	Most Representative Document begins with:
1	Disease, medicine, alternative, head, drug	New England Medical Journal in 1984 ran the heading: "Ninety Percent of Diseases are not Treatable by Drugs ...
2	austria, war, german, ally, front	World War I, also called First World War or Great War, an international conflict that in 1914–18 embroiled most ...

Table 12: Framework #1 Emerging Topics for (a) SIGIR and (b) 10Newsgroups Datasets

All detected 5 Emerging Topics for SIGIR dataset do not appear in the Background documents. Furthermore, the detected two Emerging Topics for the 10Newsgroups dataset represent both Historical and Medical Emerging Topic trends.

To verify our results, we repeated our experiment using different selections of Emerging Topics for 10Newsgroups dataset with different number of documents varies between 25 to 100 documents, and we could achieve similar results as all detected topics were always Emerging Topics, however the number of Emerging Topics varied each time because of inclusion of sub-topics that belong to Emerging Topics.

By repeating our experiment for SIGIR dataset, we noticed that the Temporal Indexing topic is sometimes replaced by another Emerging Topic that do not form a trend in the Foreground dataset. Given that there are only 3 documents in the Foreground dataset that discuss this topic, and that some of its keywords are similar to other indexing topics in the Background, the model was merging it with Background topics during some of LDA runs. We could avoid this problem by reducing the threshold of a trending topic from 3 to 2 documents, therefore the user should consider reducing the desired threshold by 1 or 2 documents when setting the hot trend’s variable.

4.5.2 Second Framework

As shown in Fig. 42, instead of representing input documents by Bag of Words, the second framework uses Term Frequency–Inverse Document Frequency (TF-IDF) representation of texts as an input for the Topic Model in the second stage. Blei & Lafferty (2009) indicated that using TF-IDF for LDA may enhance the speed of the model because it reflects in advance how significant a word is to a document in the dataset. Fig.41 shows that Coherence Scores for Topic Models with TF-IDF are generally higher than other models that use BoW representation.

To select the number of documents “K” for our TF-IDF LDA Model, Perplexity and Coherence scores are monitored while automatically increasing the model’s number of topics. At each number of topics, the number of Emerging Topics that only appear in Foreground documents is checked.

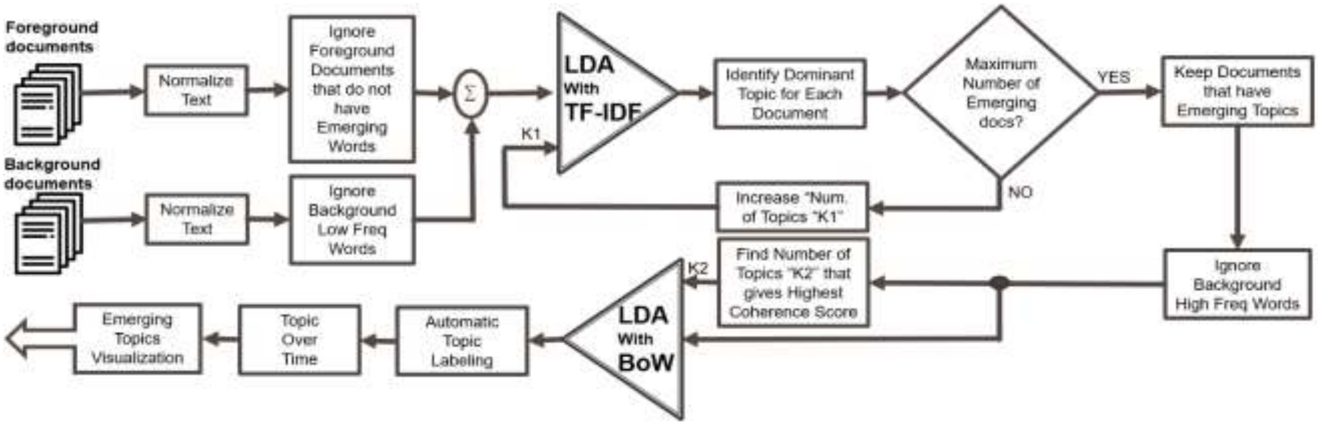

Foreground documents
Normalize Text
Ignore Foreground Documents that do not have Emerging Words
Σ
LDA With TF-IDF
Identify Dominant Topic for Each Document
Maximum Number of Emerging docs?
YES
Keep Documents that have Emerging Topics
K1
Background documents
Normalize Text
Ignore Background Low Freq Words
Increase "Num. of Topics "K1"
NO
K2
Find Number of Topics "K2" that gives Highest Coherence Score
Ignore Background High Freq Words
Emerging Topics Visualization
Topic Over Time
Automatic Topic Labeling
LDA With BoW

Figure 42: Emerging Topic Detection Filtered-LDA Framework #2

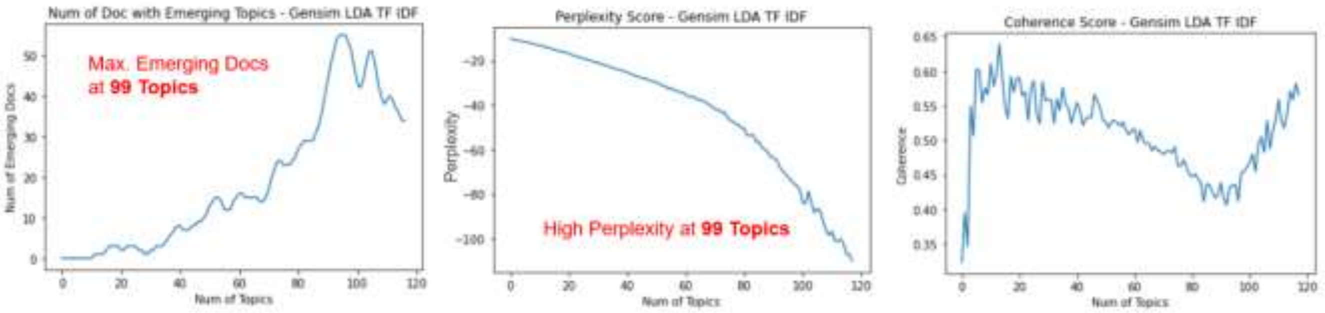
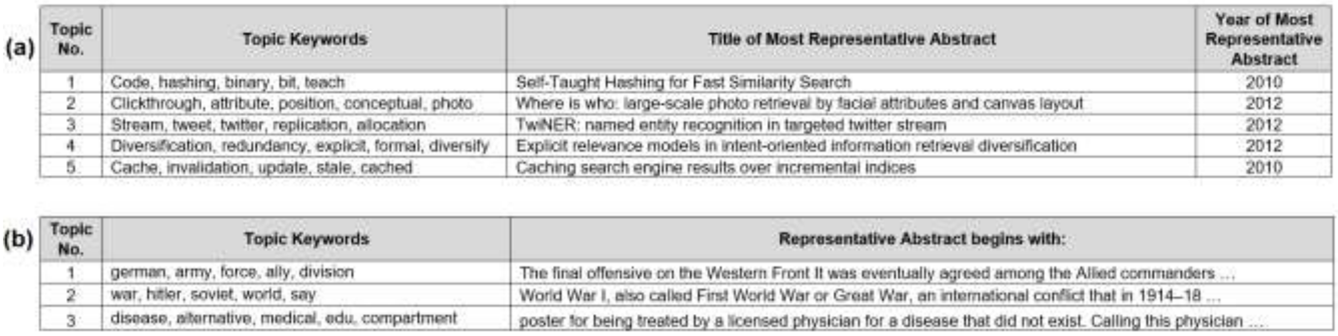


Figure 43: 10Newsgroups Perplexity & Coherence Scores for LDA with TF-IDF

Finally, the system automatically selects the number of topics that produces maximum number of documents that contain Emerging Topics because this represents the state in which the model performs best clustering. In other words, it is the state when the model represents the test data well. It is also noticed that the model has a good Perplexity score at this setpoint.

For the 10Newsgroups dataset, Fig. 43 shows the curves of Perplexity scores, Coherence Scores, and the number of Emerging Topics against the setpoint of the number of topics "K". The optimum number of topics that provides maximum number of documents with Emerging Topics is 99, whereas the number for SIGIR dataset increased to 100 topics.



(a)

Topic No.	Topic Keywords	Title of Most Representative Abstract	Year of Most Representative Abstract
1	Code, hashing, binary, bit, teach	Self-Taught Hashing for Fast Similarity Search	2010
2	Clickthrough, attribute, position, conceptual, photo	Where is who: large-scale photo retrieval by facial attributes and canvas layout	2012
3	Stream, tweet, twitter, replication, allocation	TwiNER: named entity recognition in targeted twitter stream	2012
4	Diversification, redundancy, explicit, formal, diversify	Explicit relevance models in intent-oriented information retrieval diversification	2012
5	Cache, invalidation, update, stale, cached	Caching search engine results over incremental indices	2010

(b)

Topic No.	Topic Keywords	Representative Abstract begins with:
1	german, army, force, ally, division	The final offensive on the Western Front It was eventually agreed among the Allied commanders ...
2	war, hitler, soviet, world, say	World War I, also called First World War or Great War, an international conflict that in 1914–18 ...
3	disease, alternative, medical, edu, compartment	poster for being treated by a licensed physician for a disease that did not exist. Calling this physician ...

Table 13: Framework #2 Emerging Topics for (a) SIGIR and (b) 10Newsgroups datasets

As shown in Table 13, the output of 2nd Framework for SIGIR dataset includes 5 Emerging Topics, however the Temporal Indexing topic was not detected as it was merged with a Background topic. Instead of Temporal Indexing, a fifth Emerging Topic is detected although it does not represent a trend. For the 10Newsgroups dataset, the Framework could detect both Emerging Topics successfully without adding any Background topic, however the Historical topic appears twice in the output.

4.6 RELATED WORK

As clarified earlier, our research work addresses capturing Emerging Topic when two sets of documents are compared, therefore research work related to Event Detection and Tracking is not really relevant to our work as the task of our frameworks is not detecting bursts of events only, but to detect Emerging Topics even when they are not among the top popular topics.

Erten et al. (2004) visualized temporal patterns to detect Emerging Topics using citations and relations between documents from conference proceedings of Association for Computing Machinery (ACM). Their proposed visualization method identifies gradually fading topics and fast-growing topics. It detects the highest five frequently words that are used in the papers' titles. However, it is not easy to apply this technique on other unlabeled datasets as it utilizes the ACM special classification of labeled papers to discover the research emerging topics in each year.

Le, Ho & Nakamori (2005) introduced a model to discover scientific trends from academic articles. The model employs both Hidden Markov Models (HMMs) (Rabiner, 1989) and Maximum-Entropy Markov Models (MEMMs) (McCallum et al., 2000), and it could achieve a maximum accuracy of 58% when tested on 100 papers.

The developers of FB-LDA Model applied it on the application of scientific trend detection. Their model could discover Emerging Topics when applied on SIGIR dataset. However, many old topics were also detected among the Emerging Topics of FB-LDA. Furthermore, the developers of FB-LDA did not provide clear guidelines for selecting the variable "K".

Shen & Wang (2020) used the VOSviewer software, the Lookup tool of MS Excel, and the Charts of M S PowerPoint to track topics related to the Perovskite Solar Cell (PSC). However, their approach requires manual tuning of the threshold related to minimum term's occurrences. Furthermore, Emerging Topics are visually detected by the user in this approach as it did not propose any mechanism to automatically articulate detected topics. A similar topic visualization approach was proposed by Swan & Jensen (2000), Havre et al. (2002), Van Eck & Waltman (2010), Vorontsov et al. (2015), and Fedoriaka (2016).

Some other Emerging Topic Detection methods are domain- specific, hence their performance on other domains is unknown. For example, Bolelli, Ertekin, & Giles (2009) proposed an LDA-Based model using Segmented Author Topic through inserting the temporal ordering inside CiteSeer research papers. Like PLDA, their model divides the duration of the publications into multiple time slots, it applies LDA on the first time slot, then it tries to correlate the topics of other time slots with the LDA topics of the first time slot. Morinaga & Yamanishi (2004) introduced a model using a standard clustering process to examine the variations in time of the detected components to monitor Emerging Topics. However, their method relies on an email collection of documents, which makes it difficult to evaluate their model's performance on other domain of documents, like research papers. Behpour et al. (2021) could avoid such domain-dependency limitation by introducing a temporal bias to the clustering process and they could improve the sharpness of topic trends, which means that they could enhance the model's ability to discover Emerging Topics.

Marrone (2020) used an entity linking approach to examine topic popularity in a set of Information Science Journal's papers. This approach uses a knowledge base to link word strings to entities, then it automatically identifies topics based on entity mentions. Multiple indicators are used to identify which topics are active. Although this approach could solve the problem of identifying the number of topics for standard Topic Modeling methods, it has a limitation of relying on rapidly growing topics. This makes this approach closer to the Event Detection methods which cannot identify Emerging Topics before they form a surge in the topic's pipeline.

4.7 Summary

In this chapter, we described three main base methods for Emerging Topic Detection, then we introduced two new frameworks, which could achieve better results when compared to base methods. The new frameworks could not only detect most of Emerging Topics successfully from both SIGIR and 10Newsgroups datasets, but they could also block old trends and major background topics.

The 1st framework can zoom into the text by varying the value of Alpha hyperparameter to detect trending topics, even when they appear only in few documents. Such low frequency topics could be merged with other background topics when a fixed value of Alpha is selected, or a low number of topics is specified.

The 2nd framework uses LDA with TF-IDF text representation, which showed higher Coherence and lower Perplexity scores when compared to standard LDA Models with BoW text representation. As a result, good topic clustering could be achieved without the need of applying cascaded LDA blocks and varying the value of Alpha hyperparameter. However, the computational cost of the 2nd framework is still higher than the 1st one because of the automatic search for maximum number of documents that contain a dominant Emerging Topic. Furthermore, unlike the 1st framework, the 2nd one failed to capture the Temporal Indexing topic for the SIGIR dataset, which gives an advantage to the 1st framework.

The proposed Emerging Topic Detection model is general and can be applied to other applications where two sets of documents are compared to discover new topics. We shall use the new model in Chapter 6, where Emerging Topic Detection is used to interpret sentiment variations for a large Twitter dataset.

Table 14 summarizes the results of our experiments in this chapter.

MODEL	Emerging Topic Detection for SIGIR Dataset	Emerging Topic Detection for 10Newsgroups Dataset
DTM	Not detected	Not detected
PLDA	Not applicable (Labels are not available for this dataset)	Not detected
FB-LDA	50% of detected topics are not Emerging Topics	66% of detected topics are not Emerging Topics
Filtered-LDA (1st Framework)	100% accuracy, all Emerging Topics are detected, no old topic is presented.	100% accuracy, all Emerging Topics are detected, no old topic is presented.
Filtered-LDA (2nd Framework)	80% accuracy, 4 out of 5 Emerging Topics are detected, no old topic is presented.	100% accuracy, all Emerging Topics are detected, no old topic is presented.

Table 14: Comparing results of Emerging Topic Detection Models

Chapter 5 : Sentiment Analysis for Short Texts*

5.1 Introduction

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 (Pang & Lee, 2008). The 1980's witnessed significant research work on Sentiment Analysis, like analyzing subjective and objective texts (Banfield, 1982), cognitive feature of sentiments (Winograd, 1983), building affective lexicons (Clore et al., 1987). In the 1990's, WordNet (Miller et al., 1990), Part of Speech (POS) Tagging (Brill, 1994), Parsing Trees based on Statistical Methods (Abney, 1996) (Manning & Schutze, 1999), directional interpretation (positive/negative/neutral) of a given text (Jacobs, 1992), predicting the semantic orientation of adjectives (Hatzivassiloglou & McKeown, 1997), and Fuzzy Model (Kruse et al., 1999) 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 (Esuli & Sebastiani, 2006) was published to provide a lexical resource like WordNet but dedicated for Sentiment Analysis, and Sentic Computing (Cambria & Hussain, 2012) 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 (Kumar & Sebastian, 2012) (Pang et al., 2002), However some Unsupervised Learning techniques (Turney, 2002) could also achieve good results.

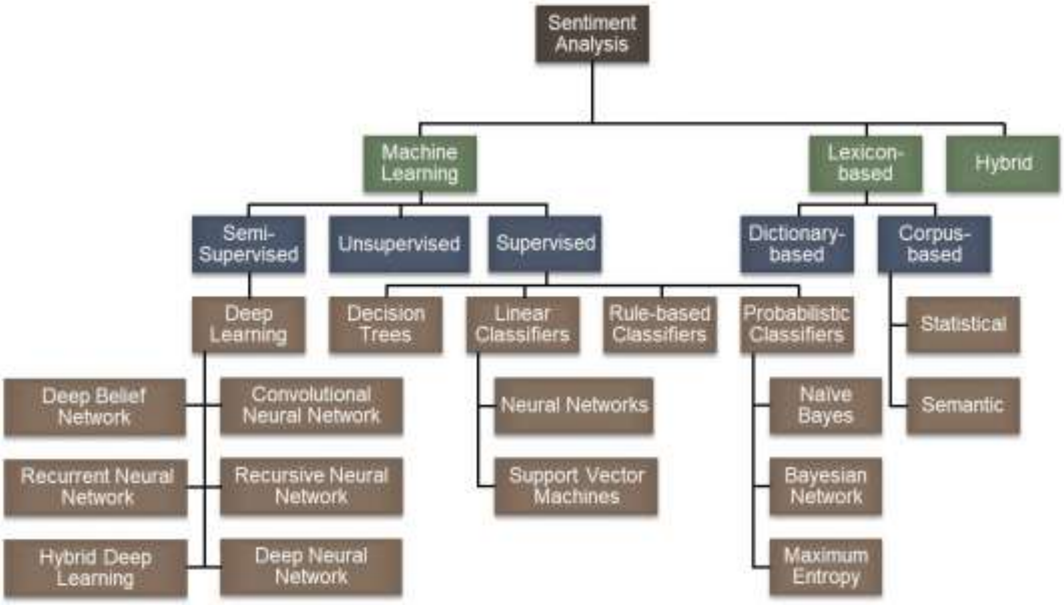


Figure 44: Main Sentiment Analysis Techniques

* This chapter is derived from the article "Alattar, F. & Shaalan, K., 2021. Using Artificial Intelligence to Understand What Causes Sentiment Changes on Social Media, Volume 9, pp. 61756-61767".

Bootstrapping method was presented to build lexicon of sentiments/subjectivity for languages that do not have enough resources (Banea et al., 2008). Semi-Supervised Learning techniques (Dalal & Zaveri, 2013) and Hybrid methods (Lu & Tsou, 2010) 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 (Dang et al., 2020), especially when Word Embedding (Mikolov et al., 2013) representation is used with the Deep Learning algorithms. The Bidirectional Encoder Representations from Transformers (BERT) algorithm (Devlin et al., 2019), which is a Neural-Network-based technique, has also shown good results when applied on Sentiment Analysis. Fig. 44 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 (Giachanou & Crestani, 2016a). Three Lexicon-based tools are still being used frequently by various researchers to extract sentiment from short texts. These are SentiStrength (Thelwall et al., 2010), TextBlob (Loria, 2020), and VADER (Hutto & Gilbert, 2014).

5.2 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 (Crowdower, 2015), which is a relatively small-size dataset that 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 (Go et al., 2009) 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.

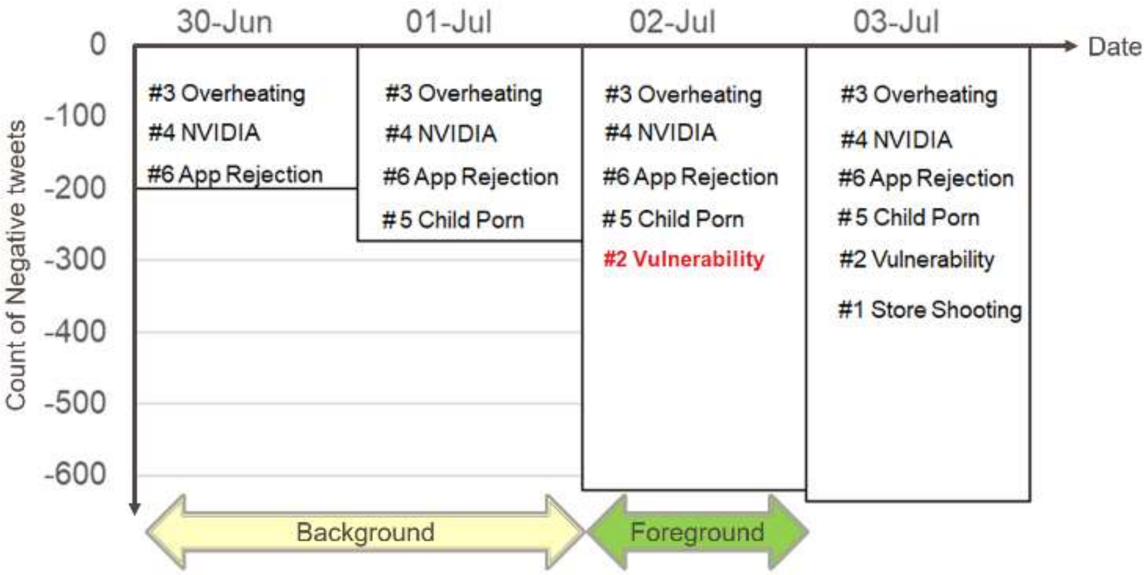


Figure 45: Ground Truth Dataset main topics

Furthermore, we are not satisfied with the quality of Sentiment140 annotation as we could easily identify many annotation errors. 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 large-size 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 (Kwak et al., 2010). To compare our Filtered-LDA results with the FB-LDA, 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 by Alattar & Shaalan (2021a) 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. 45 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 that caused major sentiment variation.

5.3 Experimental Setup

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.

We used a Dell Inspiron 7370 laptop with Intel$^{(R)}$ Core$^{(TM)}$ i7-8550U CPU @ 1.99 GHz Processor, Installed RAM is 16.0 GB, Windows 10 Home version 20H2 64-bit operating system. Python version 3.8.3 is used with Jupyter Notebook server version 6.1.4.

For text preprocessing and Sentiment Analysis classification, we used Python version 3.8.3 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 (Dang et al., 2020). It learns word embeddings by using a 2-layer Neural Network (Mikolov et al., 2013). 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 can assign sentiment levels to both Emoticons and words. For other sentiment classifiers, Emoticons are automatically replaced by their Wikipedia meanings (Wikipedia, 2021).

5.4 Comparing Sentiment Analysis Classifiers

Fig. 46 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 by Dang et al. (2020) for applying Deep Learning algorithms using Word2vec representation on US Airlines Dataset.

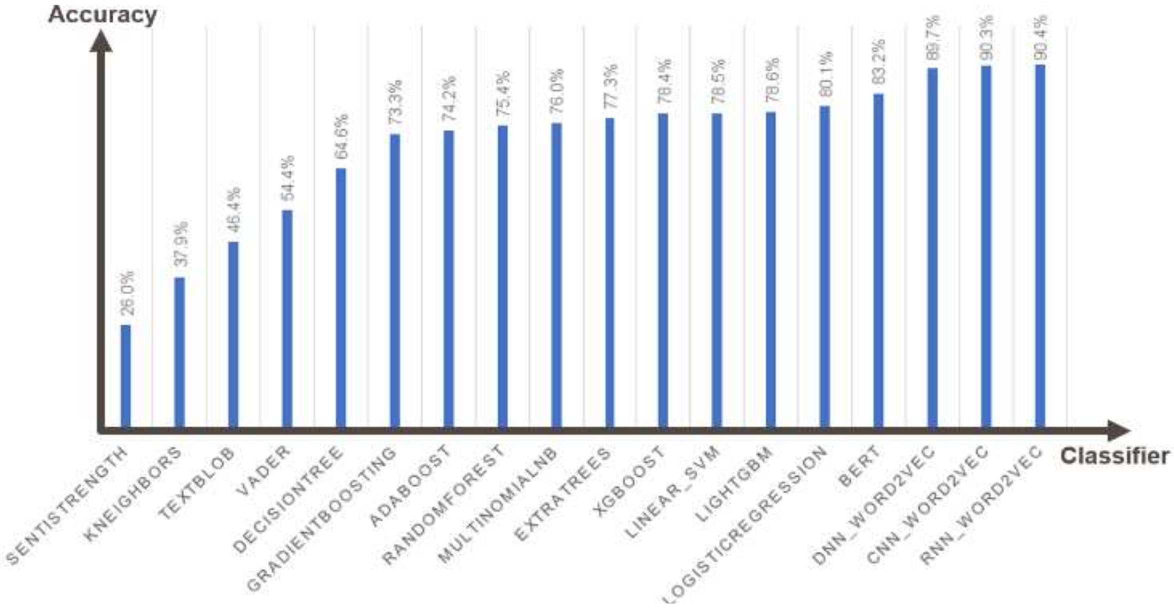


Figure 46: Accuracy of Sentiment classifiers trained and tested on same Twitter domain

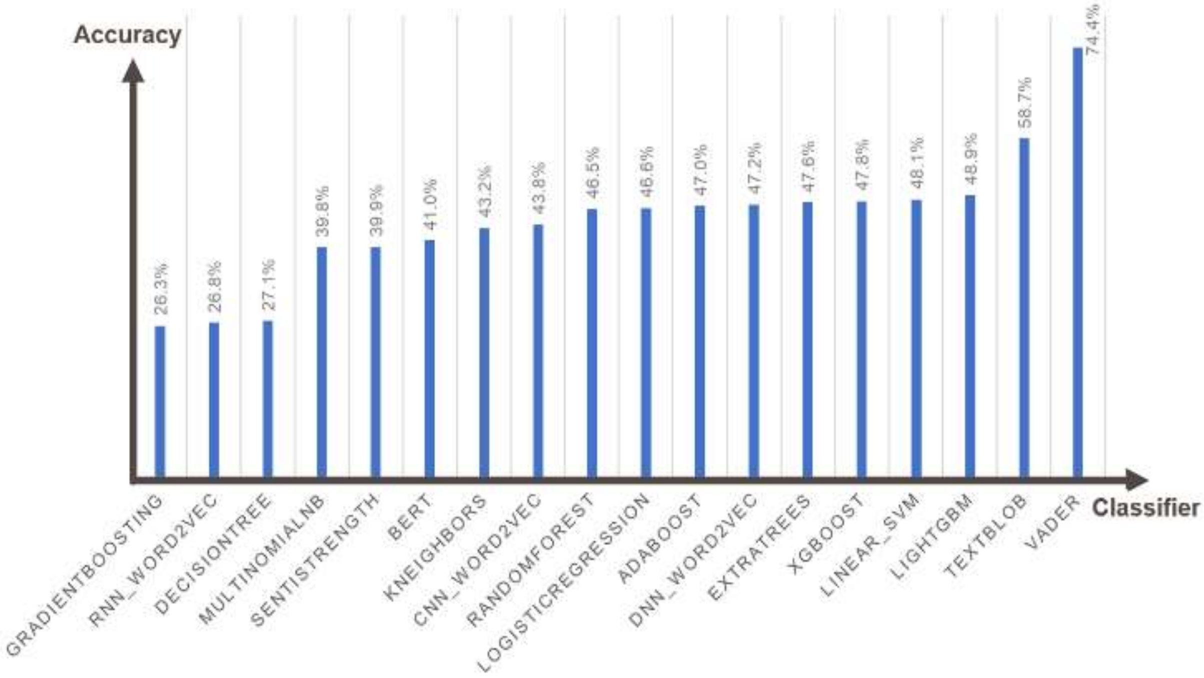


Figure 47: Accuracy of Sentiment classifiers trained on one domain and tested on a different Twitter domain

Unfortunately, all Learning-based algorithms show totally different results when they were re-tested on the Ground Truth Twitter dataset as shown in Fig. 47, 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 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. (2020) 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

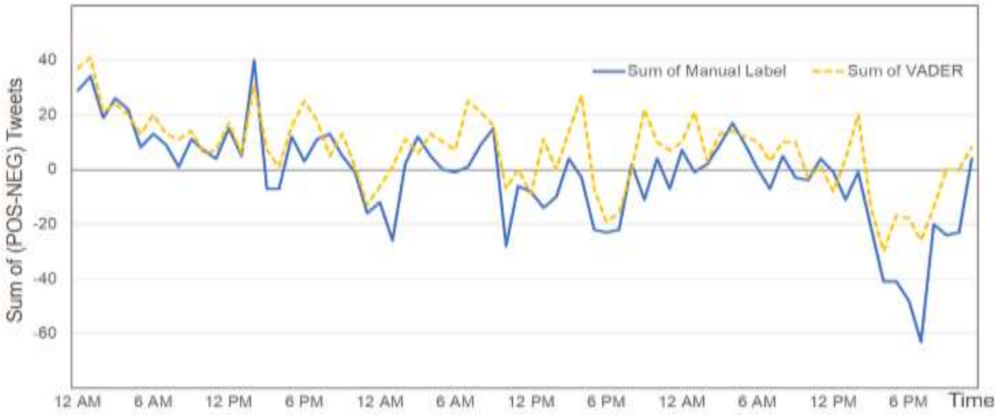


Figure 48: Hourly aggregated Overall Sentiment for Ground Truth dataset using VADER vs Manual Annotation

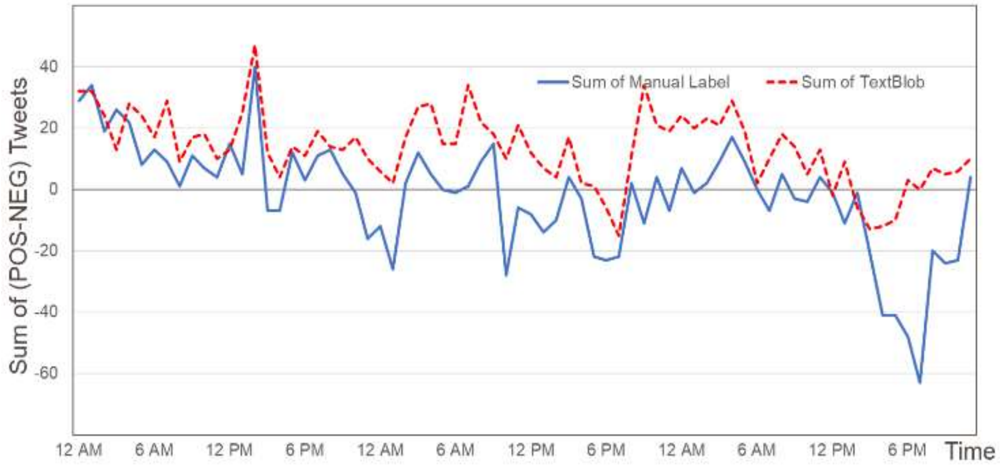


Figure 49: Hourly aggregated Overall Sentiment for Ground Truth dataset using TextBlob vs Manual

As explained by Hutto & Gilbert (2014), 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. 48 and Fig. 49 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.

5.5 Summary

In this chapter, we listed the main approaches for Sentiment Analysis, then we conducted two experiments to select the best classifier for the Stanford Twitter Dataset. In the first experiment, we trained and tested all classifiers on the US Airlines Twitter Dataset. In the second experiment, we tested the same trained classifiers of the first experiment on the Ground Truth Dataset, which is extracted from the Stanford Twitter Dataset.

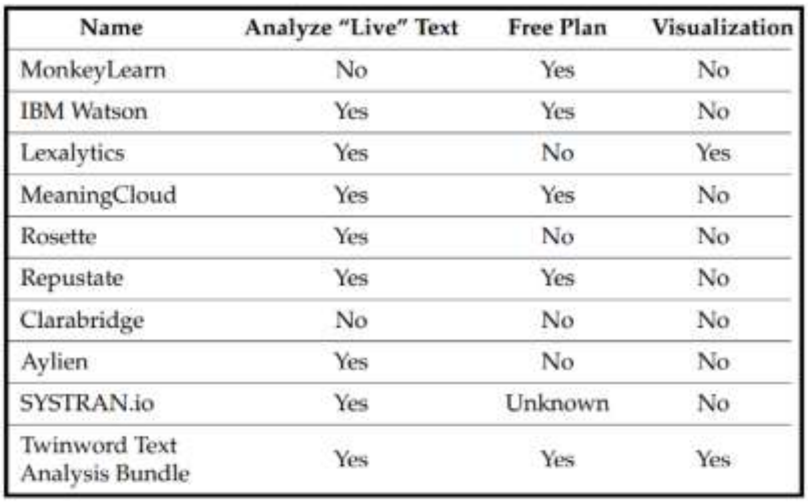
Our experiments showed that Learning-Based classifiers could not perform well when they were tested on a totally different domain. We believe that the main reason for this low performance is the relatively small size of the US Airlines Twitter Dataset and the different domain of the Ground Truth Dataset. Therefore, we decided to select VADER for our short-text Sentiment Classification because it showed highest accuracy for the Ground Truth Dataset. VADER uses a mixture of Rule-based and Lexicon-based approaches; therefore, it does not need any training. As a result, VADER's performance does not depend on the domain or the size of training datasets.

In the next chapter, we shall apply the selected Sentiment Classifier for short text, VADER, on the large Stanford Twitter Dataset, and we'll use our new Emerging Topic Detection model – which we developed in Chapter 4 – to extract the main reasons for sentiment variations.

Chapter 6 : A Framework for Sentiment Reason Mining*

6.1 Introduction

Hundreds of millions of tweets are being posted every day to discuss various topics (Tighe et al., 2015) 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 15 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.



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

Table 15: Some Sentiment Analysis software applications (Shakhovska et al., 2020)

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. (2020) attempted to interpret changes of public sentiment towards Covid-19 on Twitter. They used Dynamic Topic Model (DTM) (Blei & Lafferty, 2006) 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.

* This chapter is derived from the article "Alattar, F. & Shaalan, K., 2021. Using Artificial Intelligence to Understand What Causes Sentiment Changes on Social Media, Volume 9, pp. 61756-61767".

Some researchers decided to tackle the challenge of interpreting public sentiment and understanding its changes over time. Nevertheless, Poria et al. (2020) predicted that Sentiment Reasoning will be among the top future directions of Sentiment Analysis field. In this chapter, we focus on the problem of automatic discovery of reasons behind sentiment variations on Twitter.

6.2 Related Work

As explained earlier, Sentiment Reason Mining aims to resolve two problems: first is finding the reason for 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 (Alattar & Shaalan, 2021a). 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.

Giachanou, Mele, & Crestani (2016) used SentiStrength (Thelwall, 2010) 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 (Alattar & Shaalan, 2021a).

Moreover, this method relies on the accuracy of LDA for tracking evolution of Topics Over Time. Fig. 50 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

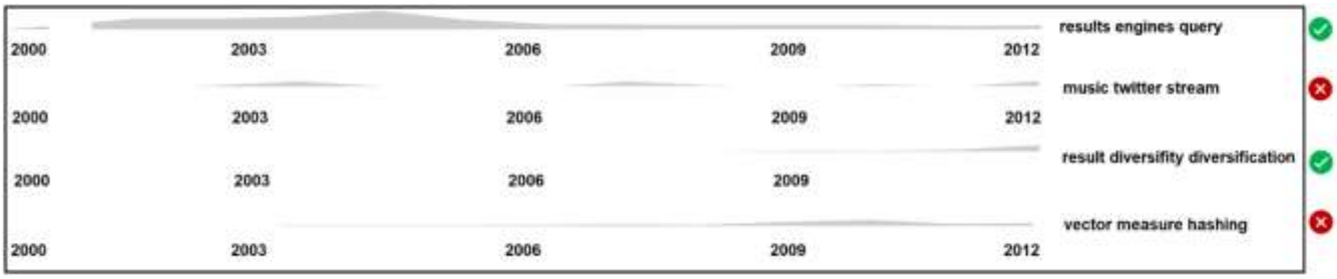
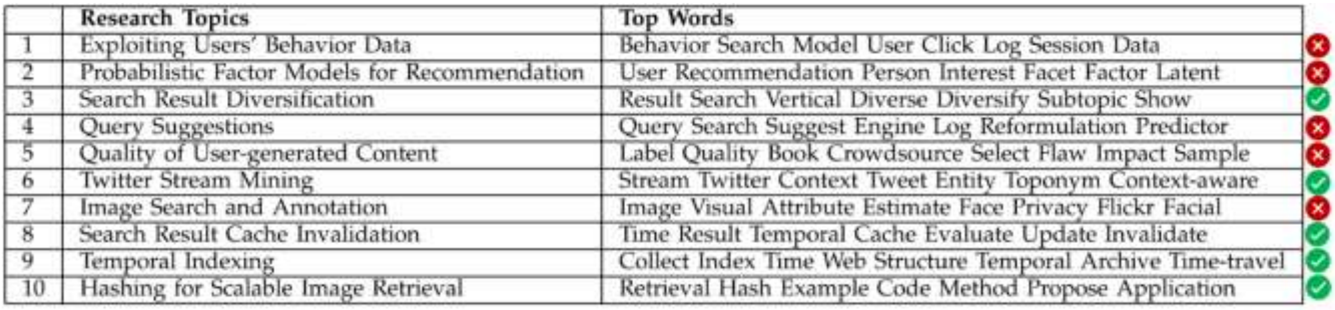


Figure 50: Topics Over Time for SIGIR dataset using dfr-Browser (Goldstone et al., 2014)



	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	✓

Table 16: FB-LDA results for SIGIR dataset (Tan et al., 2014). We added the side labels to mark the correctly detected Emerging Topics

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.

FB-LDA Model was developed by Tan et al. (2014) 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.

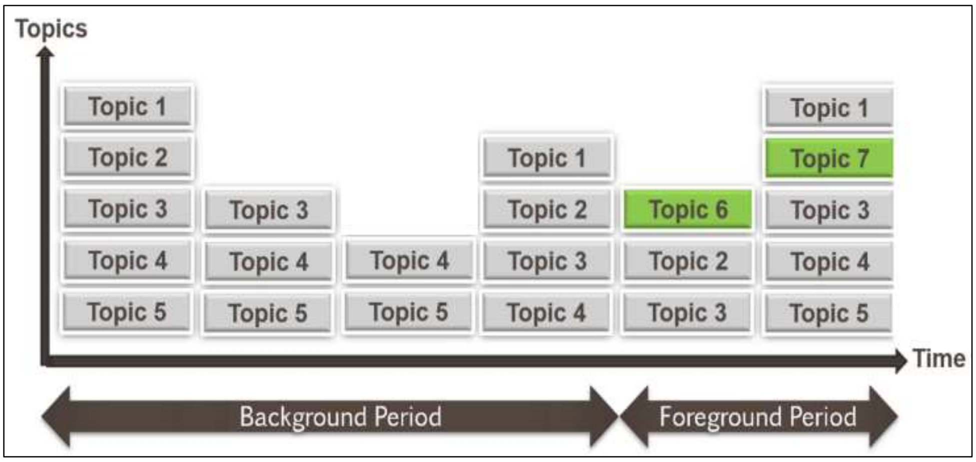


Figure 51: Emerging Topics appear only in Foreground documents

Fig. 51 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. (2014) 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 16, 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.

Event Detection method for analyzing sentiment variations was tailored by Jiang, Meng & Yu (2011) who were inspired by the Topic-Sentiment Mixture (TSM) models (Mei et al., 2007) to trace abrupt increases in document number for discussed topics, then correlate them with sentiment variations. Their framework is called Topic Sentiment Change Analysis (TSCA). It uses a rule-based method to extract sentiment, and Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 2009) 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 (Alattar& Shaalan, 2021).

6.3 A Filtered-LDA Framework for Sentiment Reason Mining

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.

Given that our study focuses only on the Sentiment Reasoning task, we shall not propose a new method for extracting sentiment level. Our experiment will use VADER for Twitter as concluded by our experiments in Chapter 5. However, our proposed Filtered-LDA framework can work with any Sentiment Analysis tool that produces acceptable sentiment classification results.

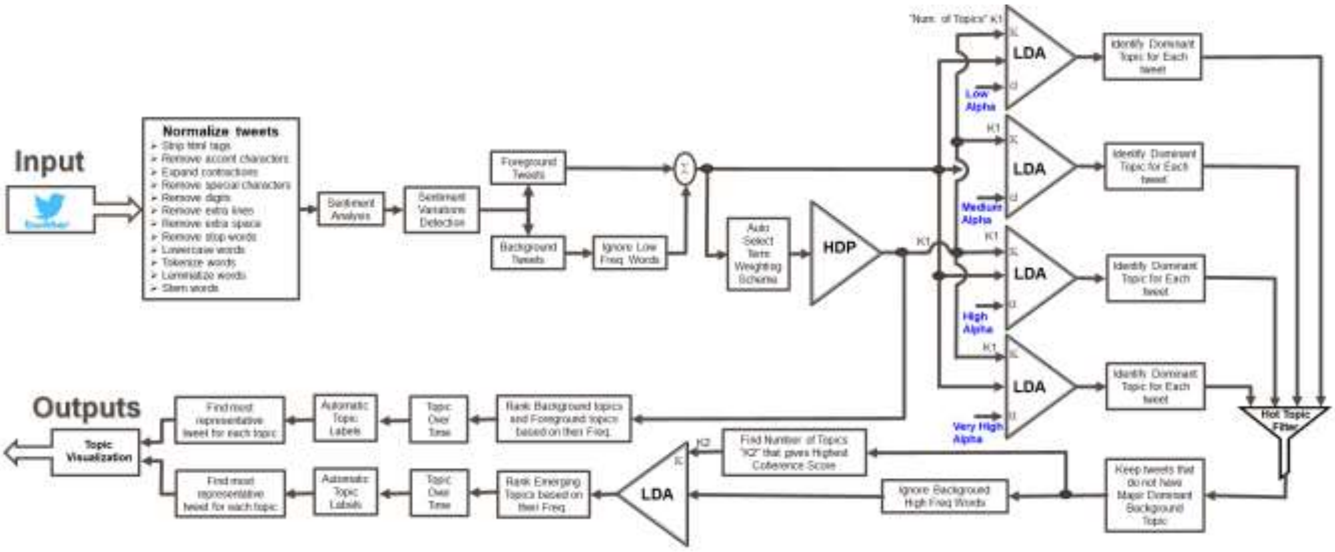

Input
Twitter
Normalize tweets
Strip html tags
Remove accent characters
Expand contractions
Remove special characters
Remove digits
Remove extra lines
Remove extra space
Remove stop words
Lowercase words
Tokenize words
Lemmatize words
Stem words
Sentiment Analysis
Sentiment Variations Detection
Foreground Tweets
Background Tweets
Ignore Low Freq. Words
Σ
Auto Select Term Weighting Schema
HDP
K1
"Num. of Topics" K1
K
LDA
α
Low Alpha
K1
LDA
Medium Alpha
K1
LDA
High Alpha
K1
LDA
Very High Alpha
Identify Dominant Topic for Each tweet
Identify Dominant Topic for Each tweet
Identify Dominant Topic for Each tweet
Identify Dominant Topic for Each tweet
Hot Topic Filter
Keep tweets that do not have Major Dominant Background Topic
Ignore Background High Freq Words
Find Number of Topics "K2" that gives Highest Coherence Score
K2
K
LDA
Rank Emerging Topics based on their Freq
Rank Background topics and Foreground topics based on their Freq.
Topic Over Time
Topic Over Time
Automatic Topic Labels
Automatic Topic Labels
Find most representative tweet for each topic
Find most representative tweet for each topic
Topic Visualization
Outputs

Figure 52: Sentiment Reason Mining Framework for Twitter

Fig. 52 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. The developers of FB-LDA 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 by Alattar & Shaalan (2021a), 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. Similar to Tan et al. (2014), 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) (Teh et al., 2004) is applied on the full cleaned dataset. Since the accuracy of HDP relies on its Term Weighting Scheme (TM) (Wilson & Chew, 2010), 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 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 (Hansen, 2016). 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 (Mei et al., 2007) 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.

6.4 Twitter Datasets

We shall use the large-size 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 (Kwak et al., 2010). We extracted all 643,264 English STD-2009 tweets that discuss "Apple", and 1,354,394 tweets that discuss "Obama".

A Ground Truth dataset was used for Apple Tweets in Chapter 4 experiments to demonstrate the high performance of Filtered-LDA when compared to FB-LDA. In the next section, we shall apply the Filtered-LDA framework on all Apple and Obama tweets in STD-2009 large dataset.

We selected this dataset so that we can compare our results with Tan et al (2014) who had used the same dataset to test their FB-LDA Model.

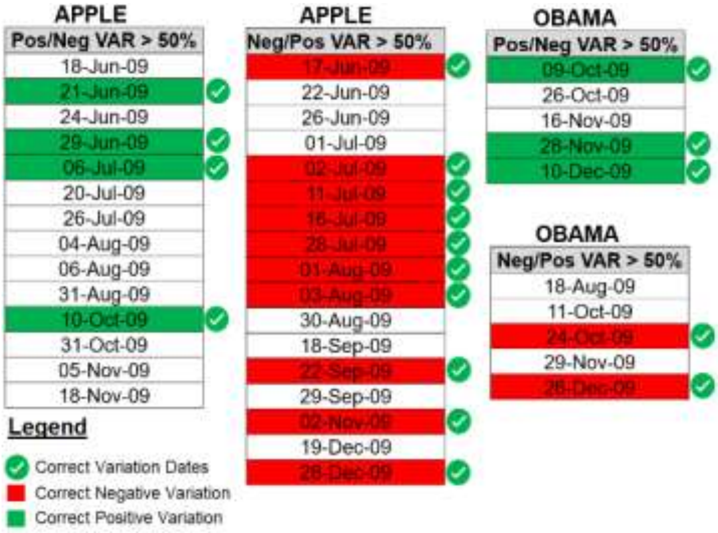
6.5 Experiments for Finding Reason Candidates

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.

6.5.1 Experimental Setup

We used a Dell Inspiron 7370 laptop with Intel(R) Core(TM) i7-8550U CPU @ 1.99 GHz Processor, Installed RAM is 16.0 GB, Windows 10 Home version 20H2 64-bit operating system. Python version 3.8.3 is used with Jupyter Notebook server version 6.1.4.

MALLET (McCallum, 2002) wrapper for Gensim (Rehurek, 2021) is used to apply LDA with optimized Gibbs Sampling (Yao et al., 2009) in our framework. We did not use standard Gensim LDA although it is faster because it employs Variational Bayes sampling method (Hoffman, Bach & Blei, 2010) which gave us lower Coherence Scores in our experiments. We used tomotopy toolkit to apply HDP with Gibb Sampling and to utilize available Automatic Topic Label module (Bab2min, 2021a). For the Cascaded LDA, four different values of Alpha are used. Low (α=1), Medium (α=50), High (α=100), and Very High (α=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.



APPLE

Pos/Neg VAR > 50%	
18-Jun-09	
🟩 21-Jun-09	✓
24-Jun-09	
🟩 29-Jun-09	✓
🟩 06-Jul-09	✓
20-Jul-09	
26-Jul-09	
04-Aug-09	
06-Aug-09	
31-Aug-09	
🟩 10-Oct-09	✓
31-Oct-09	
05-Nov-09	
18-Nov-09	

APPLE

Neg/Pos VAR > 50%	
🟥 17-Jun-09	✓
22-Jun-09	
26-Jun-09	
01-Jul-09	
🟥 02-Jul-09	✓
🟥 11-Jul-09	✓
🟥 16-Jul-09	✓
🟥 28-Jul-09	✓
🟥 01-Aug-09	✓
🟥 03-Aug-09	✓
30-Aug-09	
18-Sep-09	
🟥 22-Sep-09	✓
29-Sep-09	
🟥 02-Nov-09	✓
19-Dec-09	
🟥 28-Dec-09	✓

OBAMA

Pos/Neg VAR > 50%	
🟩 09-Oct-09	✓
26-Oct-09	
16-Nov-09	
🟩 28-Nov-09	✓
🟩 10-Dec-09	✓

OBAMA

Neg/Pos VAR > 50%	
18-Aug-09	
11-Oct-09	
🟥 24-Oct-09	✓
29-Nov-09	
🟥 28-Dec-09	✓

Legend

✓	Correct Variation Dates
🟥	Correct Negative Variation
🟩	Correct Positive Variation

Table 17: Sentiment variation dates using criteria applied by Tan et al. (2014). We marked the correct positive variation periods in green color, and the correct negtive variation periods in red color

6.5.2 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. 53 and Fig. 54 show the sentiment curves of "Apple" and "Obama" respectively.

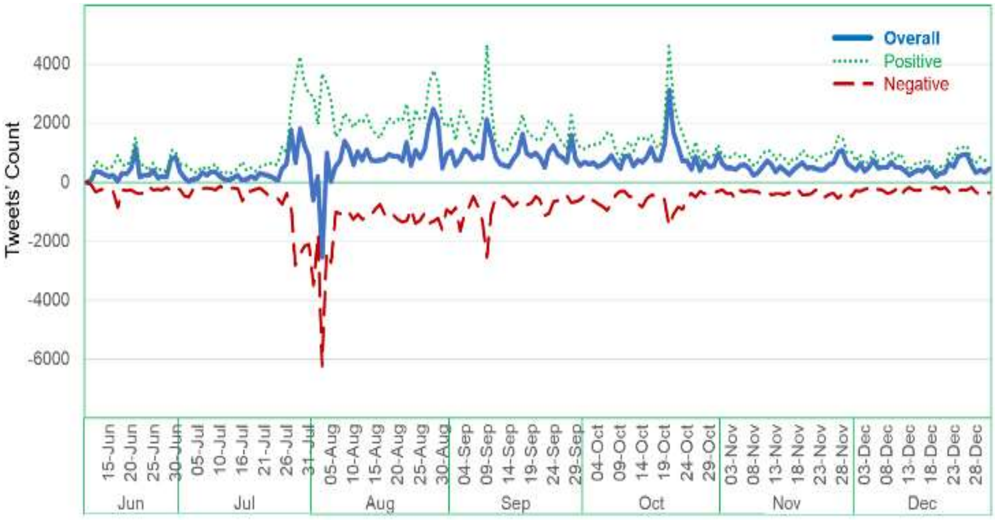


Figure 53: Daily aggregated "Apple" sentiments using VADER

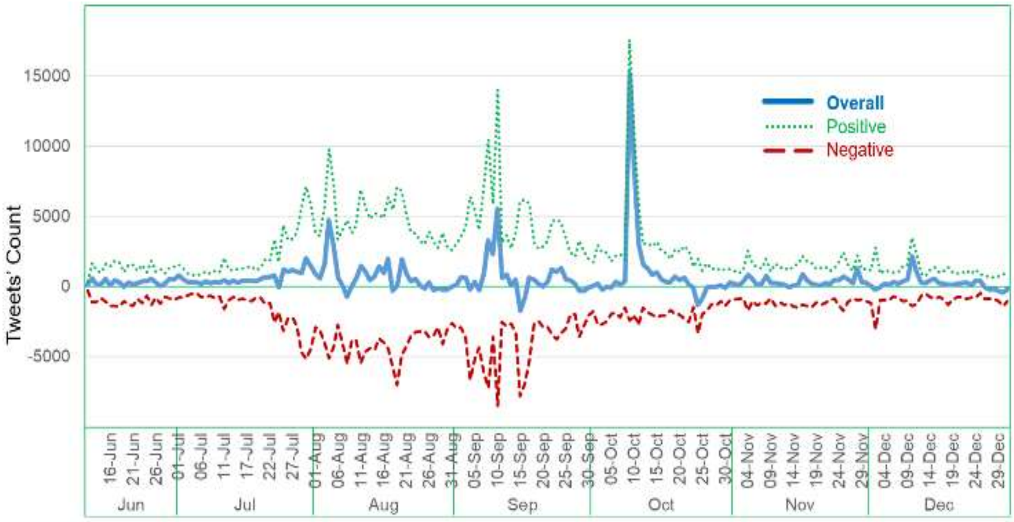
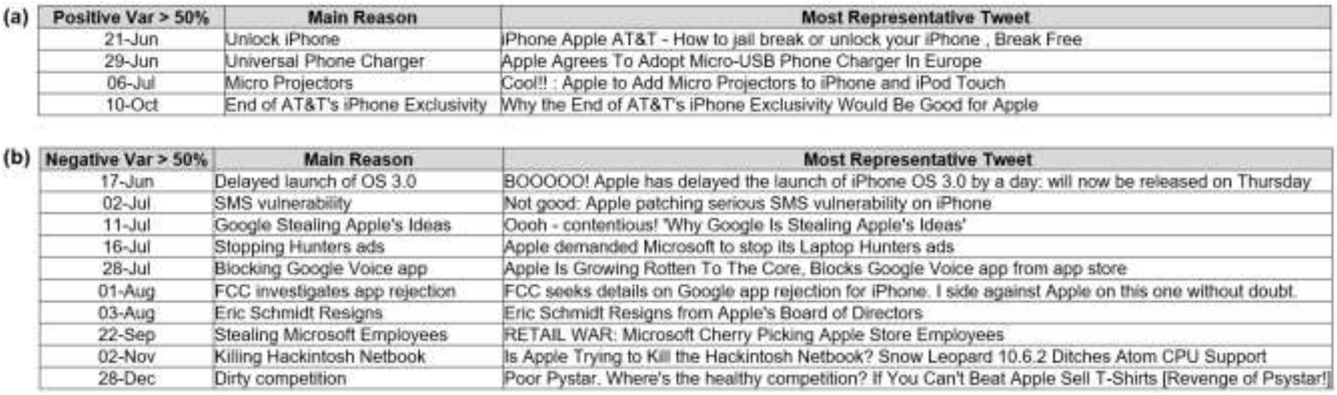


Figure 54: Daily aggregated "Obama" sentiments using VADER

As shown earlier in Fig. 52, once the Sentiment Analysis is completed, positive and negative sentiment variation periods are identified. Table 17 shows all sentiment variation dates using the measurement criteria of FB-LDA 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 by Alattar & Shaalan (2021a), where major increase in positive and negative sentiment levels are also considered in the measurement process. 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.

6.5.3 Emerging Topic Detection



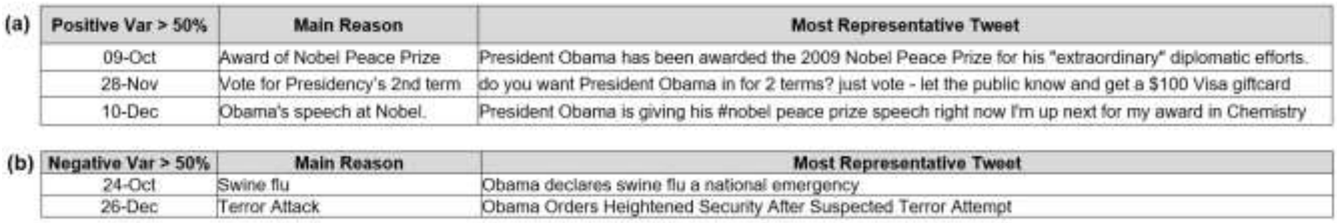
(a)

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 18: Filtered-LDA reason candidates for (a) positive and (b) negative “Apple” 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

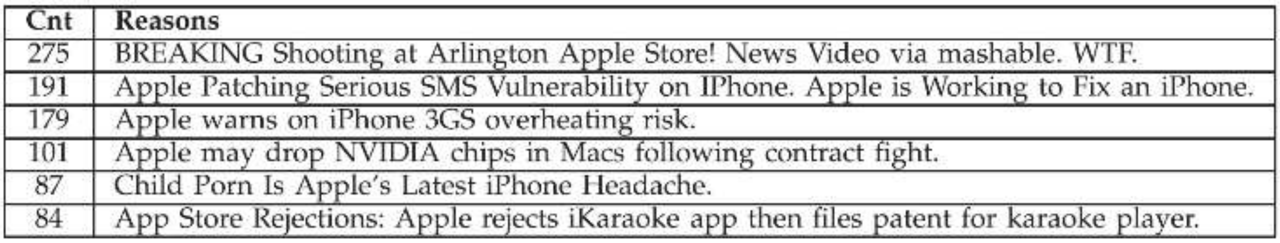
(b)

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

Table 19: Filtered-LDA reason candidates for (a) positive and (b) negative “Obama” sentiment variations

For the final LDA Model, we selected the highest frequency Emerging Topic to represent main Reason Candidate for each sentiment variation. Table 18 and Table 19 summarize the automatically detected Reason Candidates. 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 Tan et al. (2014), 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, which employs proposed TextRank algorithm's implementation method by Barrios et al. (2015), or the BERT Summarizer (K-tahiro, 2021), which employs text summarization with Pretrained Encoders (Liu & Lapata, 2019).

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.



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.

Table 20: RCB-LDA reason candidates for sentiment variation towards "Apple" from 1st to 3rd of July (Tan et al., 2014)

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, the developers of FB-LDA also introduced the Reason Candidate and Background LDA (RCB-LDA) Model to rank the candidate reasons. Table 20 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 for negative sentiment variation towards Apple from 1st to 3rd of July 2009. As clear from Fig. 45, 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 for sentiment spike is the "SMS Vulnerability", which is detected successfully by the Filtered-LDA framework as shown in Table 18-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.

6.5.4 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.

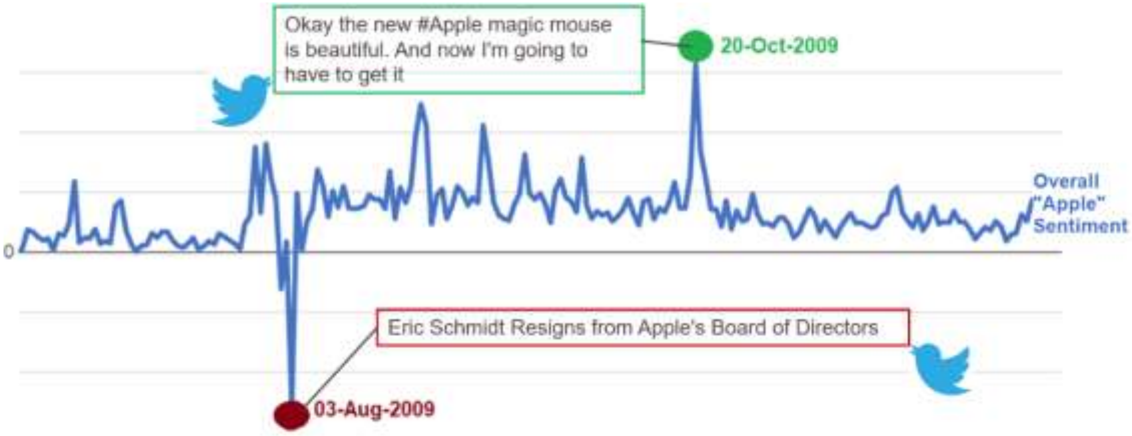


Figure 55: Overall Sentiment Spikes' Reason Candidates Visualization for "Apple"

If the sum of positive tweets is higher than negative tweets, the overall sentiment is positive, and vice versa. As shown earlier in Fig. 53 and Fig. 54, the overall sentiment levels of "Apple" and "Obama" show multiple positive and negative spikes throughout the 7 months period.

Fig. 55 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. 56 shows the same for Obama's overall sentiment. With this threshold, only 1 positive peak 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 18-b. Therefore, the most representative tweet of the topic about "Eric Schmidt Resignation" is shown in Fig. 55 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. 55 linked to the overall positive peak of 20th of October.

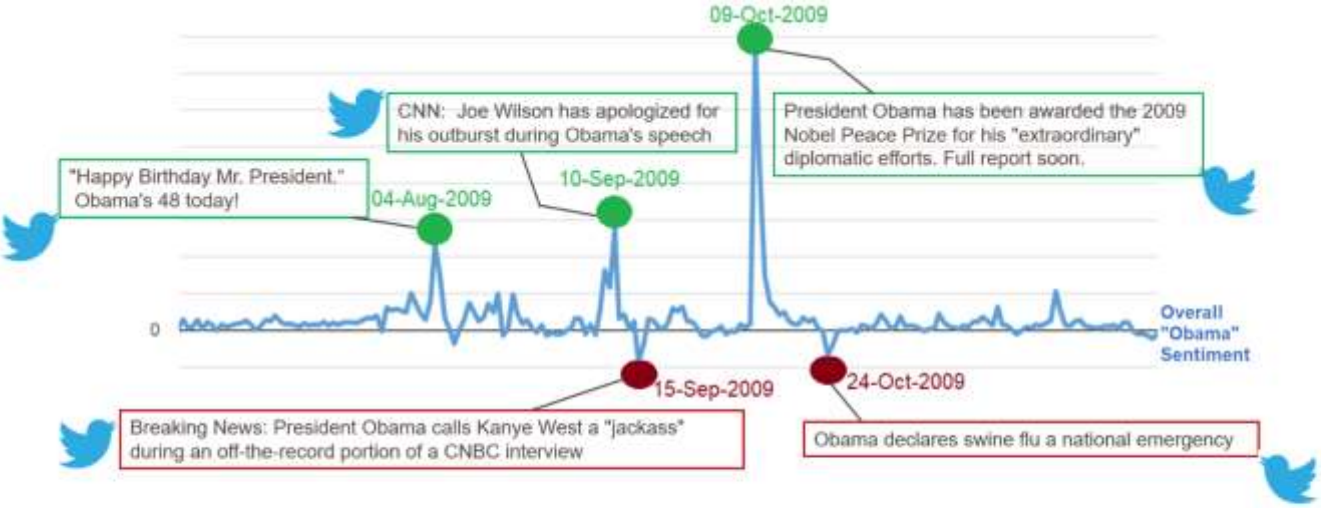


Figure 56: Overall Sentiment Spikes' Reason Candidates Visualization for "Obama".

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. 56 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.

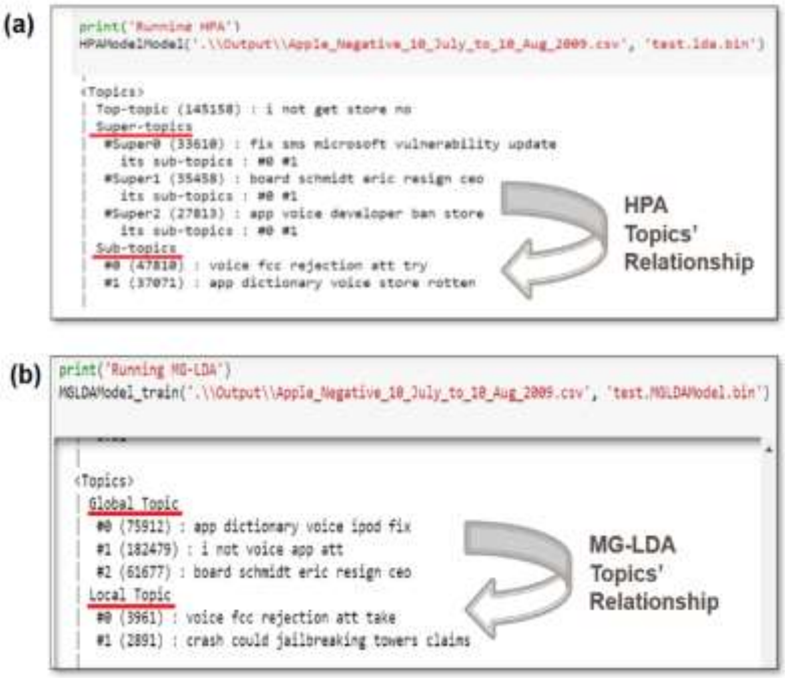
First negative peak happened when Obama mentioned the word "jackass" about the American rapper, Kanye West, during an interview. Second negative peak was caused by Obama's announcing Swine Flu a national emergency.

6.5.5 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 18-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 18-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) (Mimno et al., 2007) and Multi Grain Latent Dirichlet Allocation (MG-LDA) (Titov & McDonald, 2008) separately to investigate the relationships between Reason Candidates. For instance, Fig. 57 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 model 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 by MG-LDA. We used tomotopy toolkit's HPA and MG-LDA functions to apply both models.



```
print('Running HPA')
HPAModelModel('.\\Output\\Apple_Negative_10_July_to_10_Aug_2009.csv', 'test.lda.bin')

<Topics>
| Top-topic (145158) : i not get store no
| Super-topics
|  #Super0 (33610) : fix sms microsoft vulnerability update
|    its sub-topics : #0 #1
|  #Super1 (35458) : board schmidt eric resign ceo
|    its sub-topics : #0 #1
|  #Super2 (27813) : app voice developer ban store
|    its sub-topics : #0 #1
| Sub-topics
|  #0 (47810) : voice fcc rejection att try
|  #1 (37071) : app dictionary voice store rotten
|
```

```
print('Running MG-LDA')
MGLDAModel_train('.\\Output\\Apple_Negative_10_July_to_10_Aug_2009.csv', 'test.MGLDAModel.bin')

<Topics>
| Global Topic
|  #0 (75912) : app dictionary voice ipod fix
|  #1 (182479) : i not voice app att
|  #2 (61677) : board schmidt eric resign ceo
| Local Topic
|  #0 (3961) : voice fcc rejection att take
|  #1 (2891) : crash could jailbreaking towers claims
```

Figure 57: (a) HPA and (b) MG-LDA outputs for one month of Apple tweets from 10th July 2009

6.6 Summary

In this chapter, we used the first new Filtered-LDA framework - which we developed in Chapter 4 – to interpret sentiment variations for a real-life large Twitter dataset, the Stanford Twitter dataset. Our new framework outperformed the best-performing base method, the FB-LDA.

Our Filtered-LDA framework could detect all sentiment variation period for both “Apple” and “Obama” targets. Moreover, the main reason for each positive and negative sentiment variation has been identified accurately by our new framework, whereas the base method of FB-LDA could not detect all sentiment variation periods accurately, and it showed some old topics among the detected Emerging Topics.

Finally, we used both the Hierarchical Pachinko Allocation (HPA) and the Multi Grain Latent Dirichlet Allocation (MG-LDA) models separately to discover the relationships between candidate reasons for sentiment variations.

Chapter 7 : Conclusions

7.1 Summary

This thesis addressed the problem of automatically detecting reasons behind major sentiment variations on Twitter. It reviewed existing methods and identified their major shortcomings. To overcome those, we proposed a novel Filtered-LDA Model that could outperform base methods. The model applies an enhanced sentiment variation measurement to detect changes on sentiment levels accurately.

The main task of the Filtered-LDA Model is to detect Emerging Topics from two sets of documents. We proposed two different frameworks to build the model. The first framework uses a Cascaded LDA block with multiple LDA hyperparameter values to zoom inside and outside texts for detecting Emerging Topics. The second framework applies TF-IDF text representation for LDA to enhance its clustering performance.

Both frameworks were applied on the application of discovering hot scientific topics from a large set of abstracts, and they showed good performance when compared to base models. However, the first framework outperformed the second framework as it could capture more Emerging Topics and its program execution is faster. The frameworks can be used in any other application that compares two large sets of documents to discover new topics and common topics.

The first framework was selected to build a Sentiment Variations' Reasoning framework for Twitter. 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 visualized the Reason Candidates by linking 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.

7.2 Answers to Research Questions

Based on the research work and experiments that have been carried out in the previous chapters, we can now summarize the answers to the research questions as follows:

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

The answer is certainly yes. After analyzing multiple real-life Twitter datasets, we noticed that majority of positive and negative tweets do include the explicit reasons of the expressed sentiment. Furthermore, due to the short length of tweet's texts, they normally include one dominant topic only, which represents the main reason for the expressed sentiment/opinion most of the time. Therefore, detecting the dominant topic of subjective tweets is important for the Sentiment Reasoning task.

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

The answer is no. As shown in some real-life tweets, reasons for positive/negative sentiment are sometimes events, which cannot be categorized as features or aspects of the target itself. In such cases, Aspect-Based methods cannot capture the reason for sentiment.

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

By analyzing many real-life Twitter's sentiment spikes, we realized that the highest-frequency topic is not always the main reason for a sentiment spike. In some cases, a positive spike will occur when discussion on a negative topic dies out, and vice versa. In some other cases, the highest frequency topic stays as a dominant topic for a long period with a stable frequency that does not form a spike, however some lower-frequency topics could emerge during the same period to cause spikes.

Having said that, it is still useful to identify the highest-frequency topic inside a spike to learn the background topic that could have relationship with the main reason for the spike.

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

No, the Event Detection method cannot always capture main reasons for sentiment variations. In our experiments, we noticed that main reasons for sentiment variation may not cause sentiment spikes because they emerge when other topics are dying. In such cases, the sentiment level would stay almost stable without forming any spike.

RQ (5) Emerging Topic: Is Emerging Topic Detection efficient for interpreting sentiment variations?

Yes, Emerging Topic Detection method showed highest efficiency when compared to other methods in our Sentiment Variations' Reason Mining experiments. Nonetheless, manual review of thousands of real-life tweets also showed that major sentiment variations correlate with highest-frequency Emerging Topics. Furthermore, Emerging Topic Detection method is not only useful for identifying the reason for a sentiment variation, but also for discovering the reason for the sentiment itself during any period. As clarified in the above answer for RQ (4), sentiment level may stay stable throughout a long period even though main reason for that sentiment keeps changing during the same period. In such cases, Emerging Topic Detection method can capture all reasons for sentiment, even when they do not cause any sentiment spike.

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

Indeed - as shown in our experiments - drawing Topic Over Time curves helps in understanding the reasons for sentiments and sentiment variations. It gives a good idea on birth, evolution, and death of each topic. Other visualization methods like topic bubbles of pyLDAvis help us understand the wight of each topic and its possible similarity to other topics.

However, due to possible merger of topics by a Topic Model in some cases, topic curves may become misleading. Wrong selection of topic numbers and/or hyperparameters may cause such topic merger.

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

No, it does not always cause a spike. In our experiments, we noticed real-life Twitter cases where a positive sentiment variation nullifies a negative sentiment variation when both positive and negative reasons for sentiment appear during same period. In such cases, overall sentiment level stays stable. Therefore, it is important to track each of positive and negative sentiment levels separately, in addition to the overall sentiment level.

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

Sure, conventional Topic Modeling methods are useful for interpreting sentiment variations. As explained in the above answer of RQ (1), reasons for sentiments are expressed explicitly inside most of subjective tweets, therefore capturing dominant topics during a certain period helps us understand possible reasons for sentiment variations during that period. However, careful attention shall be given to tuning of conventional Topic Models as these could give misleading outputs when their variables or hyperparameters are wrongly selected.

High Coherence Score for a Topic Model is a good measure for correct selection of number of topics, however a user should also consider the Perplexity Score of the model to reduce possibility of merging totally different topics together.

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

Partially, yes. In our experiments, FB-LDA could detect Emerging Topics in many cases, however its output often includes old topics as well. The main reason for such wrong detection is the lack of guidelines for selecting the number of FB-LDA topics. Furthermore, in some cases, the keywords of FB-LDA Emerging Topics included a main keyword from an old topic. This makes it difficult for a user to correctly understand the model's output.

RQ (10) Efficient Topic Detection Technique: How can Emerging Topics be efficiently detected even when they are not among high-frequency topics?

Two different techniques are proposed in this thesis to achieve efficient Emerging Topic Detection. The first technique – which shows higher accuracy – depends on selecting multiple low, medium, and high values of LDA hyperparameters to scan all possible topics inside a document set. This step is followed by a voting filter which extracts documents that contain Emerging Topics.

The second technique uses the TF-IDF text representation for the LDA Model. It identifies the number of topics that produces highest number of Emerging Topics for a set of documents. This number is then used to identify documents that contain Emerging Topics.

Both techniques are followed by a final Topic Model with highest Coherence Score to identify the Dominant Emerging Topic that has the highest frequency. We developed frameworks for implementing these techniques to apply the proposed Filtered-LDA Model for Emerging Topic Detection.

RQ (11) Sentiment Classifier for Short Texts: What is the best choice so far for classifying sentiment on social media when domain of texts is unknown?

We recommend using VADER (Valence Aware Dictionary for sEntiment Reasoning) for short texts when domain of short texts is unknown. In our experiments, employing this tool which uses a mixture of Rule-based and Lexicon-based approaches, showed highest accuracy when compared to other

Lexicon-Based tools. Moreover, when Learning-Based Sentiment Analysis classifiers are trained on specific domains, but tested on totally different domains, VADER could outperform all Learning-Based methods. However, when both training and testing domains are similar, many Learning-Based classifiers outperformed Lexicon-Based tools, including VADER. In such cases, RNN, CNN, and DNN Deep Learning algorithms with Word Embedding representation showed highest performance, followed by the Bidirectional Encoder Representations from Transformers (BERT) algorithm.

RQ (12) Sentiment Reason Mining Dashboard: How to ensure that Human Machine Interface displays Reason Candidates in a comprehensible form?

For short texts, like tweets, we believe it is more convenient for a human to read a full topic representative tweet than trying to interpret topic content through reading conventional Topic Model's output of topic keywords, or even automatically generated topic labels. The most representative tweet for a topic can be simply selected by identifying the tweet in which that topic has had the highest probability. The graphical representation which we proposed in this thesis links the most representative tweet for each reason candidate to its caused sentiment variation. With this simple Human Machine Interface, it is easy to track sentiment variations online along with the most representative tweet for each variation.

7.3 Future Research Directions

In our future work, we shall apply the Filtered-LDA Model on Arabic tweets to check its performance for non-English texts. In addition to Sentiment Reasoning application for Arabic tweets, we shall examine performance of the model on Arabic Emerging Topic Detection using timestamped texts. For example, it is useful to compare all early verses of Quran during the Macci period with all later verses of Quran during the Madani period. Finding highest frequency unique topics in each period would help in understanding the nature of the scripture at each stage. Similar comparisons can also be carried out between Old Testament, New Testament, and Quran to identify their unique and common topics.

For tracking sentiment levels of tweets, we shall investigate other methods of monitoring sentiment by considering the importance of expressed opinion inside each tweet. The weight of each opinion shall depend on the number of its retweets and number of followers of the tweet's author. For instance, a tweet that is generated by a famous author with millions of followers should have higher weighting when compared to tweets or retweets of users who got few followers.

References

Abdelhaq, H., Sengstock, C. & Gertz, M. (2013). ‘Eventweet: online localized event detection from Twitter’. *The VLDB Endowment*, 6(12), pp. 1326–1329.

Abney, S. (1996). ‘Statistical methods and linguistics’. Cambridge, MA, The MIT Press, pp. 1-26.

Admane, P. et al. (2016). ‘Interpreting the public sentiment variation on social media’. *Global Journal of Advanced Engineering Technologies*, 5(1), pp. 68-71.

Alattar, F. & Shaalan, K. (2021a). ‘A survey on opinion reason mining and interpreting sentiment variations’. *IEEE Access Journal*, Volume 9, pp. 39636-39655.

Alattar, F. & Shaalan, K. (2021b). ‘Using Artificial Intelligence to Understand What Causes Sentiment Changes on Social Media’. IEEE Access Journal, Volume 9, pp. 61756-61767.

Avachar, V., Khot, P., Bhosale, O, Brahmane, P. & Jamadar, R.A. (2016). ‘implementation on interpreting the public sentiment variations on twitter’. *International Engineering Research Journal (IERJ)*, 2(3), pp. 1181–1184.

Bab2min (2021a). Module tomotopy.label. [Online] Available at: https://bab2min.github.io/tomotopy [Accessed 02 April 2021].

Bab2min (2021b). tomotopy 0.11.1 Python extension. [Online] Available at: https://pypi.org/project/tomotopy/ [Accessed 02 April 2021].

Banea, Mihalcea, E. & Wiebe, J. (2008). ‘A bootstrapping method for building subjectivity lexicons for languages with scarce resources’. Marrakech, Morocco, European Language Resources Association (ELRA), pp. 2764–2767.

Banfield, A. (1982). ‘Unspeakable sentences: narration and representation in the language of fiction’. New York, USA: Law Book Co of Australasia.

Barrios, F., Lopez, F., Argerich, L. & R. Wachenchauzer (2015). ‘Variations of the similarity function of textrank for automated summarization’. Rosario, Argentina.

Behpour, S., Mohammadi, M., Albert, M.V., Alam, Z.S., Wangc, L., & Xiao, T. (2021). "Automatic trend detection: Time-biased document clustering," Knowledge-Based Systems, vol. 220, no. 106907, pp. 1-12.

Bhalerao & Dange, T. (2015). ‘Interpretation of public sentiment variations using tweets’. *International Journal of Innovative Research in Computer and Communication Engineering,* 3(6), pp. 5828-5834.

Blair-Goldensohn, S. et al. (2008). ‘Building a sentiment summarizer for local service reviews’. Beijing; China.

Blei, D., Jordan, M., Griffiths, T. & Tenenbaum, J. (2004). ‘Hierarchical topic models and the nested chinese restaurant process’. 17–24.

Blei, D. & Lafferty, J. (2005). Correlated topic models’. pp. 147–154.

Blei, D. & Lafferty, J. (2006). ‘Dynamic topic models’. pp. 113-120.

Blei, D. & Lafferty, J. (2009). ‘Topic models’. *Text mining: classification, clustering, and applications Journal,* 15 June, 10(71), pp. 1-34.

Blei, D.M. & McAuliffe, J.D. (2010). ‘Supervised topic models’. *Advances in Neural Information Processing Systems,* pp. 1-22.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). ‘Latent dirichlet allocation’. *Journal of Machine Learning Research (JMLR),* pp. 993–1022.

Bolelli, L., Ertekin, S., & Giles, C. (2009). "Topic and Trend Detection in Text Collections Using Latent Dirichlet Allocation," in ECIR2009: 31st European Conference on Information Retrieva, Toulouse, France.

Boltuzic, F. & Snajder, J. (2014). ‘Back up your stance: recognizing arguments in online discussions’. Baltimore, Maryland USA, pp. 49–58.

Botchway, R., Jibril, A., Oplatkova, Z. & Chovancova, M. (2020). ‘Deductions from a sub-saharan african bank’s tweets: a sentiment analysis approach’. *Cogent Economics & Finance,* 8(1), pp. 1-19.

Brill, E. (1994). ‘Some advances in transformation-based part of speech tagging’. Seattle, Washington, AAAI Press, pp. 722–727.

Brody, S. & Elhadad, N. (2010). ‘An unsupervised aspect-sentiment model for online reviews’. Los Angeles, California, pp. 804–812.

Cambria, E. & Hussain, A. (2012). ‘Sentic computing techniques, tools and applications’. New York, USA: Springer.

Cambria, K., Schuller,B., Xia, Y., & Havasi, C. (2013). ‘New avenues in opinion mining and sentiment analysis’. *IEEE Intelligent Systems,* 28(2), pp. 15-21.

Chakrabarti, D. & Punera, K. (2011). ‘Event summarization using tweets’. Barcelona, Catalonia, Spain.

Choi, Y. & Cardie, C. (2010). ‘Hierarchical sequential learning for extracting opinions and their attributes’. Uppsala, Sweden, pp. 269–274.

Chen, G., Kong, Q. & Mao, W. (2017). ‘Online event detection and tracking in social media based on neural similarity metric learning’. Beijing, China, pp. 182-184.

Chen, P., Chen, S. & Liu, J. (2020). ‘Hierarchical sequence labeling model for aspect sentiment triplet extraction’. Zhengzhou’, China.

Chen, Y., Yin, C. Y., Lin, Y. J. & Zuo, W. (2018). ‘On-line evolutionary sentiment topic analysis’ *Modeling. International Journal of Computational Intelligence Systems,* Volume 11, pp. 634–651.

Chiru, C., Rebedea, T. & Ciotec, S. (2014). ‘Comparison between LSA-LDA lexical chains’. Barcelona, Spain, pp. 255–262.

Clore, G., Ortony, A. & Foss, M. (1987). ‘The psychological foundations of the affective lexicon’. *Journal of Personality and Social Psychology,* 53(4), pp. 751-766.

ComArg (2014). ‘Corpus of Online User Comments with Arguments’. [Online] Available at: http://takelab.fer.hr/data/comarg [Accessed 16 April 2021].

Crowdower (2015). ‘Twitter US Airline Sentiment’. [Online] Available at: https://www.kaggle.com/crowdflower/twitter-airline-sentiment [Accessed 30 Jan 2021].

Cvitanic, T. et al. (2016). ‘LDA v. LSA: A comparison of two computational text analysis tools for the functional categorization of patents’. Atlanta, GA, USA.

D'Andrea, A., Ferri, F., Grifoni, P. & Guzzo, T. (2015). ‘Approaches; tools and applications for sentiment analysis implementation’. *International Journal of Computer Applications*, 125(3), pp. 26–33.

Dalal, M. & Zaveri, M. (2013). ‘Semisupervised learning based opinion summarization and classification for online product reviews’. *Applied Computational Intelligence and Soft Computing,* Volume 2013, pp. 1-8.

Dang, N.C., Moreno-Garcia, M.N. & Prieta, F.D. (2020). ‘Sentiment analysis based on deep learning: a comparative study’. *Electronics Journal,* 9(483), pp. 1-29.

Deerwester, S. et al. (1990). ‘Indexing by latent semantic analysis’. *Journal of the American Society for Information Science,* 41(6), pp. 391–407.

Devlin, J., Chang, M., Lee, K. & Toutanova, K. (2019). ‘BERT: pre-training of deep bidirectional transformers for language understanding’. Minneapolis, Minnesota, pp. 4171–4186.

Dhanush, D., Thakur, A. K. & Diwakar, N. P. (2016). ‘Aspect-based sentiment summarization with deep neural networks’. *International Journal of Engineering Research & Technology (IJERT),* 5(5), pp. 371–375.

Eguchi, K. & Lavrenko, V. (2006). ‘Sentiment retrieval using generative models’. Sydney, Australia, pp. pp. 345–354.

Eguchi, K. & Shah, C. (2006). ‘Opinion retrieval experiments using generative models: Experiments for the TREC 2006 blog track’. Gaithersburg; Maryland.

Erten, C., Harding, P., Kobourov, S., Wampler, K., & Yee, G. (2004) "Exploring the computing literature using temporal graph visualization," SPIE - The International Society for Optical Engineering, vol. 5259, no. 2004, p. 45–56.

Esuli, A. & Sebastiani, F. (2006). ‘SentiWordNet: a publicly available lexical resource for opinion mining’. Genova, Italy, pp. 417-422.

Fedoriaka, D. S. (2016). ‘Hierarchic topic models visualization’. *MIPT,* pp. 1-8.

Giachanou, A. & Crestani, F. (2016a). ‘Like it or not: a survey of twitter sentiment analysis methods’. *ACM Computing Surveys,* 49(2), pp. 28-41.

Giachanou, A. & Crestani, F. (2016b). ‘Tracking sentiment by time series analysis’. Pisa Italy.

Giachanou, A., Mele, I. & Crestani, F. (2016). ‘Explaining sentiment spikes in twitter’. Indianapolis Indiana USA, Association for Computing Machinery.

Go, A., Bhayani, R. & Huang, L. (2009). 'Twitter sentiment classification using distant supervision', Stanford.

Goldstone, A. et al. (2014). 'An interactive topic model of signs'. *Signs* Journal.

Guzman, E. & Maalej, W. (2014). 'How do users like this feature? A fine-grained sentiment analysis of app reviews'. Karlskrona, Sweden, *IEEE,* pp. 153–162.

Hansen, J. (2016). 'Inside latent dirichlet allocation: an empirical exploration'. *Knowledge and Information Systems,* pp. 1-21.

Hasan, K.S. & Ng, V. (2014). 'Why are you taking this stance? identifying and classifying reasons in ideological debates'. Doha, Qatar.

Hatzivassiloglou, V. & McKeown, K. (1997). 'Predicting the semantic orientation of adjectives'. Stroudsburg, PA, USA, Association for Computational Linguistics, pp. 174–181.

Havre, S., Hetzler, E., Whitney, pp. & Nowell, L. (2002). 'Theme River: visualizing thematic changes in large document collections'. *IEEE Transactions on Visualization and Computer Graphics*, 8(1), pp. 9–20.

Hettiarachchi, H., Adedoyin-Olowe, M., Bhogal, J. & Gaber, M. M. (2020). 'Embed2Detect: temporally clustered embedded words for event detection in social media'. *Journal of Machine Learning,* pp. 1-33.

Hoffman, M., Bach, F. & Blei, D. (2010). 'Online learning for latent dirichlet allocation'. Vancouver, British Columbia, Canada, pp. 856-864.

Hofmann, M. & Chisholm, A. (2016). 'Text mining and visualization case studies using open-source tools'. New York: CRC Press.

Hofmann, T. (1999). 'Probabilistic latent semantic analysis'. Berkeley California USA, pp. 50–57.

Hu, M. & Liu, B. (2004). 'Mining and summarizing customer reviews'. Seattle, Washington, USA.

Hurst, F. & Nigam, K. (2004). 'Retrieving topical sentiments from online document collections'. *Document Recognition and Retrieval XI,* Volume 5296, pp. 27–34.

Hutto, C. & Gilbert, E. (2014). 'VADER: a parsimonious rule-based model for sentiment analysis of social media text'. Oxford, Oxfordshire, England, *Association for the Advancement of Artificial Intelligence AAAI,* pp. 1-10.

Hu, Y., John, A., Wang, F. & Seligmann, D. D. (2012). 'ET-LDA: joint topic modeling for aligning events and their twitter feedback'. Vancouver, BC, Canada.

Ingule, D. & Chhajed, G. (2014). 'Survey of public sentiment interpretation on twitter'. *International Journal of Engineering Research and General Science,* 2(6), pp. 943-947.

Ishaq, A., Asghar, S. & Gillani, S. A. (2020). 'Aspect-based sentiment analysis using a hybridized approach based on CNN and GA'. *IEEE Access,* Volume 8, pp. 135499–135512.

J. Hansen, 2016. 'Inside latent dirichlet allocation: an empirical exploration'. *Knowledge and Information Systems,* pp. 1-21.

Jacobs, P., ed. (1992). 'Direction-based text interpretation as an information access refinement'. In: *Text-based intelligent systems: current research and practice in information extraction and retrieval*. Berkeley, California, UAE: Lawrence Erlbaum Associates, pp. 257–274.

Jamadar, R.A., Avachar, V., Bhosale, O., Khot, P. & Bramhane, P. (2016). 'Survey on interpreting public sentiment variations on twitter'. *Journal of Emerging Technologies and Innovative Research* (JETIR), 3(5), pp. 287-289.

Jebbara, S. & Cimiano, P. (2016). 'Aspect-based sentiment analysis using a two-step neural network architecture'. Heraklion, Crete, Greece, Springer International Publishing, pp. 153–167.

Jeevitha, R. (2016). 'Interpreting sentimental analysis for customer commands on e-commerce'. *International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST),* 2(10), pp. 928-929.

Jiang, L. et al. (2011). 'Target-dependent twitter sentiment classification'. Portland, Oregon, pp. 151–160.

Jiang, Y., Meng, W. & C. Yu (2011). 'Topic sentiment change analysis'. Berlin, Germany, Springer, pp. 443-457.

Jin, W., Ho, H. H. & Srihari, R. K. (2009). 'OpinionMiner: a novel machine learning system for web opinion mining and extraction', Paris France.

Kamini, A. & Ezhillarasi, B. (2015). ‘Interpretation of sentiment variations on micro blogs with wall filtering’. *International Journal of Advanced Research Trends in Engineering and Technology (IJARTET),* 2(8), pp. 201-205.

Khalid, S., Aslam, M. H. & Khan, M. T. (2018). ‘Opinion reason mining: implicit aspects beyond implying aspects’. Riyadh, Saudi Arabia.

Khalid, S. & Khan, M. T. (2018). ‘A novel opinion reason mining framework exploiting linguistic associations’. Zurich, Switzerland, pp. 6–10.

Kim, S.O. & Hovy, E. (2006). ‘Automatic identification of pro and con reasons in online reviews’. Sydney, Australia, pp. 483–490.

Kruse, R., Nauck, D. & Borgelt, C. (1999). ‘Data mining with fuzzy methods: status and perspectives’. Aachen, Germany, Aachen : Verlag Mainz, pp. 1-8.

K-tahiro, 2021. ‘BERT-summarizer 0.1.4’. [Online] Available at: https://pypi.org/project/bert-summarizer/ [Accessed 02 April 2021].

Kumar, A. & Sebastian, T. (2012). ‘Sentiment analysis: a perspective on its past, present and future’. *International Journal of Intelligent Systems and Applications*, 4(10), pp. 1-14.

Kwak, H., Lee, C., Park, H. & Moon, S. (2010). ‘What is Twitter, a social network or a news media?’. Raleigh North Carolina USA, pp. 591–600.

Kwak, H., Lee, C., Park, H. & Moon, S. (2010). ‘What is twitter; a social network or a news media’. Raleigh North Carolina USA, Association for Computing Machinery, pp. 591–600.

Lang, K. (2008). ‘10Newsgroups Dataset’. [Online] Available at: https://www.kaggle.com [Accessed 21 Feb 2021].

Lee, D. D. & Seung, H. S. (1999). ‘Learning the parts of objects by non-negative matrix factorization’. *Nature,* 401(6755), pp. 788–791.

Lee, J. et al. (2018). ‘Ensemble modeling for sustainable technology transfer’. *Sustainability Journal*, 10(7), pp. 1-15.

Lee, M. & Song, M. (2020). ‘Incorporating citation impact into analysis of research trends’. *Scientometrics*, 124(2), pp. 191–1224.

Le, M., Ho, T. & Y. Nakamori (2005). ‘Detecting emerging trends from scientific corpora’. *International Journal of Knowledge and Systems Sciences,* June, 2 (2), pp. 1-7.

Leskovec, J., Backstrom, L. & Kleinberg, J. (2009). ‘Meme-tracking and the dynamics of the news cycle’. Paris France.

Li, F., Huang, M. & Zhu, X. (2010). ‘Sentiment analysis with global topics and local dependency’. Atlanta, Georgia, pp. 1371–1376.

Limsettho, N., Hata, H. & Matsumoto, K. (2014). ‘Comparing hierarchical dirichlet process with latent dirichlet allocation in bug report multiclass classification’. Las Vegas, NV, USA, pp. 1-6.

Lin, W. H., Wilson, T., Wiebe, J. & Hauptmann, A. (2006). ‘Which side are you on?: Identifying perspectives at the document and sentence levels’. New York, USA, pp. 109–116.

Liu, B. (2010). ‘Sentiment Analysis and Subjectivity’. In: *Handbook of Natural Language Processing*. London/ Boca: Chapman & Hall CRC, pp. 627–666.

Liu, P., Joty, S. & Meng, H. (2015). ‘Fine-grained opinion mining with recurrent neural networks and word embedding’. Lisbon, Portugal.

Liu, Y. & Lapata, M. (2019). ‘Text summarization with pretrained encoders’. Hong Kong, China, Association for Computational Linguistics, pp. 3730–3740.

Li, W. & McCallum, A. (2006). ‘Pachinko allocation: Dag-structured mixture models of topic correlations’. Pittsburgh, PA, pp. 577–584.

Loria, S. (2020). ‘TextBlob: Simplified Text Processing’. [Online] Available at: https://textblob.readthedocs.io/ [Accessed 26 March 2021].

Lu, B., Ott, M., Cardie, C. & Tsou, B. K. (2011). ‘Multi-aspect sentiment analysis with topic models’. Vancouver, Canada, pp. 81–88.

Lu, B. & Tsou, B. K. (2010). ‘Combining a large sentiment lexicon and machine learning for subjectivity classification’. Qingdao, China, pp. 3311-3316.

Lu, Y., Zhai, C. & Sundaresan, N. (2009). ‘Rated aspect summarization of short comments’. Madrid Spain.

Manikandan, R. & Kalpana, A. (2016). ‘Tracking and analyzing the public opinions by interpretation techniques’. *International Journal of Innovative Research in Technology,* 2(11), pp. 316 - 318.

Manning, C. & Schutze, H. (1999). ‘Foundations of statistical natural language processing’. Cambridge, MA: The MIT Press.

Marrone, M. (2020). "Application of entity linking to identify research fronts and trends," Scientometrics, vol. 122, no. 2020, pp. 357-379.

McCallum, A. (2002). ‘MAchine Learning for LanguagE Toolkit website’. [Online] Available at: http://mallet.cs.umass.edu/ [Accessed 02 April 2021].

McCallum, A., Freitag, D. & Pereira, F. (2000). ‘Maximum entropy markov models for information extraction’. 591–598.

Mei, Q., Shen, X. & Zhai, C. (2007). ‘Automatic labeling of multinomial topic models’. pp. 490-499.

Mei, Q. et al. (2007). ‘Topic sentiment mixture: modeling facets and opinions in weblogs’. Banff, Alberta, Canada, pp. 171–180.

Meng, X. et al. (2012). ‘Entity-centric topic-oriented opinion summarization in twitter’. Beijing China.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). ‘Efficient estimation of word representations in vector space’. Scottsdale, Arizona.

Miles, D. (2017). ‘A taxonomy of research gaps: identifying and defining the seven research gaps’. Dallas, Texas, pp. 1-10.

Miller, G. et al. (1990). ‘Introduction to WordNet: an on-line lexical database’. *International Journal of Lexicography,* 3(4), pp. 235-244.

Mimno, D., Li, W. & McCallum, A. (2007). ‘Mixtures of hierarchical topics with pachinko allocation’. Corvalis Oregon USA, Association for Computing Machinery, pp. 633-640.

Mimno, D. & McCallum, A. (2008). ‘Topic models conditioned on arbitrary features with Dirichlet-multinomial regression’. Helsinki, Finland, pp. 411–418.

Moghaddam, S. & Ester, M. (2010). ‘Opinion digger: an unsupervised opinion miner from unstructured product reviews’. Toronto, Ontario, Canada, pp. 1825–1828.

More, P. & Ghotkar, A. (2016). ‘A study of different approaches to aspect-based opinion mining’. *International Journal of Computer Applications*, 145(6), pp. 11–15.

Morinaga, S. & Yamanishi, K. (2004). "Tracking dynamics of topic trends using a finite mixture model," in the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, USA.

Muller-Bloch, C. & Kranz, J. (2015). ‘A framework for rigorously identifying research gaps in qualitative literature reviews’. Fort Worth, pp. 1-19.

O’Connor, B., Balasubramanyan, R., Routledge, B. R. & A, S. N. (2010). ‘From tweets to polls: linking text sentiment to public opinion time series’. Washington, DC, USA, pp. 122–129.

Ortega, R. et al. (2014). ‘UO UA using latent semantic analysis to build a domain-dependent sentiment resource’. Dublin, Ireland, pp. 773–778.

Pang, B. & Lee, L. (2008). ‘Opinion mining and sentiment analysis.’ *Foundations and Trends in Information Retrieval*, 2(1), pp. 1–135.

Pang, Lee, L. & Vaithyanathan, S. (2002). ‘Thumbs up? sentiment classification using machine learning techniques’. Philadelphia, USA, Association for Computational Linguistics, pp. 79–86.

Patil, A., Sedamkar, R. R. & Gupta, S. (2015). ‘A novel reason mining algorithm to analyze public sentiment variations on twitter and facebook’. *International Journal of Applied Information Systems (IJAIS),* 10(3), pp. 23–29.

Patil, S. & Kulkarni, S. (2018). ‘Mining social media data for understanding students learning experiences using memetic algorithm’. *Materials Today,* 5(1), pp. 693–699.

Peng, M. et al. (2018). ‘Emerging topic detection from microblog streams based on emerging pattern mining’. Nanjing, China.

Persing, I. & Ng, V. (2009). ‘Semi-supervised cause identification from aviation safety reports’. Singapore.

Poonam, W. & Kinikar, M. (2014). ‘Interpreting the public sentiment with emotions on twitter’. *International Journal for Research in Applied Science & Engineering Technology (IJRASET),* 2(12)*,* pp. 2-6.

Popescu, A. & Etzioni, O. (2005). ‘Extracting product features and opinions from reviews’. Vancouver, B.C., Canada.

Poria, S., Cambria, E. & Gelbukh, A. (2016). ‘Aspect extraction for opinion mining with a deep convolutional neural network’. *Knowledge-Based Systems,* 108, 15 September 2016, Pages 42-49, Volume 108*,* pp. 42-49.

Poria, S., Hazarika, D., Majumder, N. & Mihalcea, R. (2020). ‘Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research’. *IEEE Transactions on Affective Computing,* pp. 1-26.

Qui, G., Liu, B., Bu, J. & Chen, C. (2011). ‘Opinion word expansion and target extraction through double propagation’. journal computational linguistics, 37(1), pp. 9–2.

Rabiner, L. (1989). ‘A tutorial on hidden markov models and selected applications in speech recognition’. *Proceedings of the IEEE, February,* 77(2)*,* pp. 257-286.

Ramage, D., Hall, D., Nallapati, R. & Manning, C. D. (2009). ‘Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora’. Singapore, pp. 248–256.

Ramage, D., Manning, C. & Dumais, S. (2011). ‘Partially labeled topic models for interpretable text mining’, pp. 457-465.

Rehurek, R. (2021). ‘Gensim 4.0.1 Python library for topic modelling’. [Online] Available at: https://pypi.org/project/gensim/ [Accessed 02 April 2021].

Rizvana, N., Kabeer, A., Gan, K. H. & Haris, E. (2018). ‘Domain-specific aspect-sentiment pair extraction using rules and compound noun lexicon for customer reviews’. *Informatics,* 5(4), pp. 1-28.

Roder, M., Both, A., & Hinneburg, A. "Exploring the Space of Topic Coherence Measures," in WSDM '15: the Eighth ACM International Conference on Web Search and Data Mining, Shanghai China, 2015.

Sakaki, T., Okazaki, M. & Matsu, Y. (2010). ‘Earthquake shakes twitter users: Real-time event detection by social sensors’. Raleigh, North Carolina, USA, pp. 851–860.

Saraiva, F., Silva, T. d. & Macedo, J. d. (2020).’ Aspect term extraction using deep learning model with minimal feature engineering’. Grenoble, France, Advanced Information Systems Engineering, pp. 185-198.

Scaffidi, C. et al. (2007). ‘Red Opal: product-feature scoring from reviews’. San Diego California USA, Association for Computing Machinery.

Shakhovska, K., Shakhovska, N. & Vesely, P. (2020). ‘The sentiment analysis model of services providers feedback’. *Electronics Journal*, 9(11), pp. 1-15.

Sharma, A. K. (2020). ‘Understanding latent dirichlet allocation (LDA)’. [Online] Available at: https://www.mygreatlearning.com/ [Accessed 30 January 2021].

Shen, X. & Wang, L. (2020). ‘Topic evolution and emerging topic analysis based on open source software’. *Journal of Data and Information Science*, 5(4), pp. 126–136.

Sievert, C. & Shirley, K. (2014). ‘LDAvis: A method for visualizing and interpreting topics’. Baltimore, Maryland, USA, pp. 63–70.

Smatana, M., Martinkova, V., Marsalekova, D. & Butka, P. (2019). ‘Interactive tool for visualization of topic models’. *Acta Electrotechnica et Informatica*, 19(2), pp. 45–50.

Swan, R. & Jensen, D. (2000). ‘TimeMines: Constructing timelines with statistical models of word usage’. Boston, MA, pp. 73–80.

Syamala, M. &, Nalini, N.J. (2019). ‘A deep analysis on aspect based sentiment text classification approaches’. *The International Journal of Advanced Trends in Computer Science and Engineering*, 8(5), pp. 1795–1801.

Tan, S. (2018). ‘FB-LDA Python implementation of the collapsed gibbs sampler for foreground and background latent dirichlet allocation.’ [Online] Available at: https://github.com/laos1984/FB-LDA

Tan, S., Li, Y., Sun, H., Guan, Z., Yan, X., Bu, J. Chen, C. & He, X. (2014). ‘Interpreting the public sentiment variations on twitter’. *IEEE Transactions on Knowledge and Data Engineering*, 26(5), pp. 1158 - 1170.

Tao, J. & Fang, X. (2020). ‘Toward multi-label sentiment analysis: a transfer learning based approach’. *Journal of Big Data,* 7(1), pp. 1-26.

Teh, Y. W., Jordan, M. I., Beal, M. J. & Blei, D. M. (2004). ‘Sharing clusters among related groups: hierarchical dirichlet processes’. Vancouver, MIT Press, pp. 1385–1392.

Thelwall, M. et al. (2010). ‘Sentiment strength detection in short informal text’. *Journal of the American Society for Information Science and Technology*, 61(12), pp. 2544-2558.

Tighe, P. et al. (2015). ‘The painful tweet: text, sentiment, and community structure analyses of tweets pertaining to pain’. *Journal of Medical Internet Research,* 17(4), pp. 1-19.

Titov, I. & McDonald, R. (2008). ‘Modeling online reviews with multi-grain topic models’. Beijing, China.

Turney, P. (2002). ‘Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews’. Philadelphia, , USA, pp. 417-424.

Urega, R. & Devapriya, M. (2015). ‘Public sentiment interpretation on twitter - a survey’. *International Journal of Engineering and Computer Science,* 4(8), pp. 13910-13916.

Van Eck, N.J & Waltman, L. (2010). ‘Software survey: VOSviewer, a computer program for bibliometric mapping’. *Scientometrics,* 84(2), pp. 523–538.

Vorontsov, K. et al. (2015). ‘BigARTM: open-source library for regularized multimodal topic modeling of large collections’. In: *Analysis of Images; Social networks and Texts*. Springer International Publications, pp. 370–384.

Wang, Q. & Ren, J. (2020). ‘Sequence prediction model for aspect-level sentiment classification’. Santiago de Compostela, Spain.

Wang, R. et al. (2015). ‘ASEM: mining aspects and sentiment of events from microblog’. Melbourne, Australia, pp. 1923–1926.

Weng, J. & Lee, B. (2011). ‘Event detection in twitter’. Barcelona, Catalonia, Spain, pp. 401–408.

Wikipedia (2021) ‘List of emoticons’. [Online] Available at: https://en.wikipedia.org/wiki/List_of_emoticons [Accessed 1 April 2021].

Wilson, A. & Chew, P. (2010). ‘Term weighting schemes for latent dirichlet allocation’. Los Angeles, California, pp. 465-473.

Wilson, T., Wiebe, J. & Hoffmann, P. (2009). ‘Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis’. *Computational Linguistics*, 35(3), pp. 399–433.

Winograd, T. (1983). ‘Language as a cognitive process: syntax’. Michigan, USA: Addison Wesley Publishing Company.

Wu, Y., Zhang, Q., Huang, X. & Wu, L. (2009). ‘Phrase dependency parsing for opinion mining’. Singapore.

Xu, J., Chen, D., Qiu, X. & Huang, X. (2016). ‘Cached long short-term memory neural networks for document-level sentiment classification’. Austin, Texas, pp. 1660–1669.

Yang, S., Rosenfeld, J. & Makutonin, J. (2018). ‘Aspect-based financial sentiment analysis using deep learning’. Lyon France, pp. 1961–1966.

Yao, L., Mimno, D. & McCallum, A. (2009). ‘Efficient methods for topic model inference on streaming document collections’. Paris France.

Yarnguy, T. & Kanarkard, W. (2018). ‘Tuning latent dirichlet allocation parameters using ant colony optimization’. *Journal of Telecommunication, Electronic and Computer Engineering*, 10(1-9), pp. 21-24.

Yessenalina, A., Choi, Y. & Cardie, C. (2010). ‘Automatically generating annotator rationales to improve sentiment classification’. Uppsala, Sweden, pp. 336–341.

Yin, H., Yang, S. & Li, J. (2020). ‘Detecting topic and sentiment dynamics due to covid-19 pandemic using social media’. Foshan, China, pp. 610-623.

Zaidan, O., Eisner, J. & Piatko, C. (2007). ‘Using annotator rationales to improve machine learning for text categorization’. Rochester, New York.

Zhang, L. & Liu, B. (2014). ‘Aspect and entity extraction for opinion mining. in: data mining and knowledge discovery for big data, methodologies, challenge and opportunities’. Berlin: Springer-Verlag, pp. 1–40.

Zhang, Q. & Lu, R. (2019). ‘A multi-attention network for aspect-level sentiment analysis’. *Future Internet*, 11(7), pp. 157-170.

Zhou, D., Chen, L. & He, Y. (2015).' An unsupervised framework of exploring events on twitter: filtering; extraction and categorization'. Austin, Texas, USA, pp. 2468-2474.

Zhuang, F. et al. (2021). 'A comprehensive survey on transfer learning'. *Proceedings of the IEEE*, 109(1), pp. 43–76.

Zhuang, L., Jing, F. & Zhu, X.-Y. (2006).' Movie review mining and summarization'. Arlington Virginia USA.

Zohuri, B. & Moghaddam, M. (2020). "Deep learning limitations and flaws'. *Modern Approaches on Material Science Journal*, 2(3), pp. 241–250.