



**NOVA**

**IMS**

Information  
Management  
School

# MAAA

---

**Mestrado em Métodos Analíticos Avançados**  
Master Program in Advanced Analytics

**Analyzing User Reviews of Messaging Apps**  
for Competitive Analysis

Wenyi Liang

Dissertation presented as partial requirement for obtaining  
the Master's degree in Data Science and Advanced Analytics

NOVA Information Management School  
Instituto Superior de Estatística e Gestão de Informação  
Universidade Nova de Lisboa

**NOVA Information Management School**  
**Instituto Superior de Estatística e Gestão de Informação**  
Universidade Nova de Lisboa

**ANALYZING USER REVIEWS OF MESSAGING APPS  
FOR COMPETITIVE ANALYSIS**

by

Wenyi Liang

Dissertation presented as partial requirement for obtaining the Master's degree in Data Science and  
Advanced Analytics

**Advisor:** Prof. Mauro Castelli

August 2021

## **DEDICATION**

This dissertation is dedicated to my parents for their constant understanding, encouragement, and support. This work is also dedicated to many friends who have enlightened me with their dreams, creativity, and bravery.

## ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor Prof. Mauro Castelli for his affirmation of the research topic, supportive encouragement, prompt feedback, and patient guidance in completing this dissertation.

I would also like to thank the faculty in the program of Data Science and Advanced Analytics. The valuable knowledge and skills they had taught me laid the academic foundation for this work. My special thanks go to Prof. Fernando Peres and Prof. Nuno Alpalhão who gave me extra lessons and guidance on homework and projects.

I wish to acknowledge the assistance provided by the staff of library and documentation services as well as academic services of NOVAIMS in finding some research materials and solving the issue of thesis registration.

I would like to extend my thanks to my classmates who had ever helped me with the study during the first academic year. They had inspired me with practical knowledge and interesting thoughts. Advices provided by Ernesto on reading research papers were really helpful.

In the end, my appreciation goes to my former supervisors in work who come from different countries but coincidentally assigned me the work of competitive analysis for peer companies or mobile apps. This work experience, to some extent, contributed to the research topic of this dissertation.

## **ABSTRACT**

The rise of various messaging apps has resulted in intensively fierce competition, and the era of Web 2.0 enables business managers to gain competitive intelligence from user-generated content (UGC). Text-mining UGC for competitive intelligence has been drawing great interest of researchers. However, relevant studies mostly focus on industries such as hospitality and products, and few studies applied such techniques to effectively perform competitive analysis for messaging apps. Here, we conducted a competitive analysis based on topic modeling and sentiment analysis by text-mining 27,479 user reviews of four iOS messaging apps, namely Messenger, WhatsApp, Signal and Telegram. The results show that the performance of topic modeling and sentiment analysis is encouraging, and that a combination of the extracted app aspect-based topics and the adjusted sentiment scores can effectively reveal meaningful competitive insights into user concerns, competitive strengths and weaknesses as well as changes of user sentiments over time. We anticipate that this study will not only advance the existing literature on competitive analysis using text mining techniques for messaging apps but also help existing players and new entrants in the market to sharpen their competitive edge by better understanding their user needs and the industry trends.

## **KEYWORDS**

Competitive analysis; Topic modeling; Sentiment analysis; Text mining; User reviews; Messaging apps

# INDEX

<b>1</b>	<b>Introduction .....</b>	<b>1</b>
<b>2</b>	<b>Related works.....</b>	<b>3</b>
2.1	Competitive analysis by text-mining UGC.....	3
2.2	Topic modeling.....	4
2.3	Sentiment analysis .....	8
<b>3</b>	<b>Methodology.....</b>	<b>11</b>
3.1	Data collection .....	12
3.2	Data preprocessing .....	13
3.2.1	Contraction expanding.....	13
3.2.2	Text cleaning .....	13
3.2.3	Word correction and normalization .....	13
3.2.3.1	Spell check.....	13
3.2.3.2	British-American spelling normalization .....	14
3.2.3.3	Spelling correction and abbreviation expansion .....	14
3.2.4	POS tagging and lemmatization .....	17
3.2.5	Non-English reviews filtering .....	18
3.2.6	Feature extraction.....	18
3.2.7	Customized stop word removal .....	19
3.2.8	Review pruning .....	19
3.3	Topic modeling.....	20
3.3.1	NMF topic model.....	20
3.3.2	Evaluation of extracted topics .....	21
3.4	Sentiment analysis .....	21
3.4.1	VADER compound score .....	21
3.4.2	Weighted sentiment score.....	21
3.4.3	Evaluation of sentiment analysis .....	22
3.5	Competitive analysis .....	23
3.5.1	Review distributions and average sentiment scores .....	23
3.5.1.1	Visual review distribution .....	24
3.5.1.2	Visual comparison of average sentiment scores .....	24
3.5.2	Sentiment evolution.....	24

<b>4</b>	<b>Results and discussion.....</b>	<b>25</b>
4.1	Results .....	25
4.1.1	Topic modeling .....	25
4.1.1.1	Extracted topics.....	25
4.1.1.2	Evaluation result of extracted topics .....	29
4.1.2	Evaluation result of sentiment analysis .....	30
4.1.3	Results of competitive analysis .....	32
4.1.3.1	Review distributions and average sentiment scores .....	32
4.1.3.2	Visual review distribution .....	33
4.1.3.3	Visual comparison of average sentiment scores .....	36
4.1.3.4	Sentiment evolution.....	37
4.2	Discussion .....	44
4.2.1	Topic modeling .....	44
4.2.2	Sentiment analysis .....	46
4.2.3	Competitive analysis .....	46
<b>5</b>	<b>Conclusion .....</b>	<b>50</b>
<b>6</b>	<b>Limitations and recommendations for future works.....</b>	<b>51</b>
<b>7</b>	<b>Bibliography .....</b>	<b>52</b>
	<b>Appendix A. Sampled reviews with wrong topics assigned by the NMF topic model .....</b>	<b>59</b>
	<b>Appendix B. Sampled reviews with inconsistent labels of sentiment polarity on weighted sentiment scores .....</b>	<b>61</b>

## LIST OF FIGURES

Figure 1: The singular value decomposition of LSA with k topics .....	5
Figure 2: Illustration of the aspect model .....	6
Figure 3: Illustration of the latent Dirichlet allocation .....	6
Figure 4: Illustration of the non-negative matrix factorization .....	7
Figure 5: The research framework of competitive analysis using topic modeling and sentiment analysis .....	11
Figure 6: Non-negative matrix decomposition on feature terms .....	20
Figure 7: Topics extracted from the NMF model .....	25
Figure 8: Review distribution by topic of Messenger .....	34
Figure 9: Review distribution by topic of WhatsApp .....	35
Figure 10: Review distribution by topic of Signal .....	35
Figure 11: Review distribution by topic of Telegram .....	36
Figure 12: Comparison of average sentiment scores by topic .....	37
Figure 13: Sentiment evolution by topic .....	38



## LIST OF TABLES

Table 1: Information of app and user reviews .....	12
Table 2: Sample of dataset with selected fields.....	12
Table 3: Additional British-American spellings .....	14
Table 4: Manual spelling corrections .....	15
Table 5: Abbreviation expansions .....	17
Table 6: POS tags for feature extraction .....	18
Table 7: Removal process of customized stop words .....	19
Table 8: Confusion matrix for sentiment evaluation .....	22
Table 9: Explanation of TP, FP, TN and FN .....	23
Table 10: Accuracies of topic extraction .....	30
Table 11: Confusion matrix of sentiment evaluation on weighted sentiment scores.....	31
Table 12: Confusion matrix of sentiment evaluation on user ratings.....	31
Table 13: Confusion matrix of sentiment evaluation on normalized VADER compound scores .....	31
Table 14: Performance of sentiment analysis.....	32
Table 15: Summary of review distributions and average sentiment scores.....	33
Table 16: Statistical aggregation of average sentiment scores by month, topic and app .....	43
Table 17: Comparison of keywords in extracted topics.....	45

## LIST OF ABBREVIATIONS AND ACRONYMS

<b>ANEW</b>	Affective Norms for English Words
<b>fLDA</b>	matrix factorization through LDA
<b>hLDA</b>	Hierarchical latent Dirichlet allocation
<b>LDA</b>	Latent Dirichlet allocation
<b>LDMM</b>	latent Dirichlet mixture model
<b>LIWC</b>	Linguistic Inquiry and Word Count
<b>LSA</b>	Latent semantic analysis
<b>LSI</b>	Latent semantic indexing
<b>MPQA</b>	Multi-Perspective Question Answering
<b>NLTK</b>	Natural Language Toolkit
<b>NMF</b>	Non-negative matrix factorization
<b>NNDSVD</b>	Non-Negative Double Singular Value Decomposition
<b>pLSA</b>	Probabilistic latent semantic analysis
<b>pLSI</b>	Probabilistic latent semantic indexing
<b>POS</b>	Part-of-speech
<b>SAFE</b>	Simple Approach for Feature Extraction
<b>SO-CAL</b>	Semantic Orientation CALculator
<b>TF-IDF</b>	Term frequency–inverse document frequency
<b>UGC</b>	User-generated content
<b>VADER</b>	Valence Aware Dictionary for sEntiment Reasoning

# 1 INTRODUCTION

The rise of various messaging apps entering the market has brought on increasingly stiff competition and understanding the real demands of users becomes vital for the success of such mobile apps. Reviews posted by users may reveal valuable intelligence for practitioners in the market. Competitive analysis by text-mining user-generated content (UGC) has been widely studied for many industries such as hospitality (He et al., 2013; G emar & Jim enez-Quintero, 2015; Amadio & Procaccino, 2016; Kim & Jeong, 2018; Gao et al., 2018; Hu & Trivedi, 2020) and retail business (Xu et al., 2011; Wang et al., 2018; Liu et al., 2019). Analyzing user reviews of messaging apps could further broaden the applications of competitive analysis leveraging text mining techniques.

Recent work has shown that competitive intelligence can be obtained through text-mining user reviews of mobile apps (Li et al., 2017; Shah et al., 2019). However, the current approaches either examine only a small part of user reviews with comparative patterns (Li et al., 2017) or requires manual filter to effectively extract app features from user reviews (Shah et al., 2019). Moreover, topic modeling was employed to extract latent topics from user reviews of similar apps, but the usefulness of the extracted topics in competitive analysis was not explored (Su et al., 2019). In general, competitive analysis using user reviews of mobile apps involves three critical issues: (1) extracting shared app aspects from informative reviews across multiple competitive apps, (2) app aspect-based sentiment analysis, and (3) competitive analysis based on app aspects and user sentiments. To the best of our knowledge, previous studies mainly focus on solving only a part of these issues. Additionally, previous methods were presented for general mobile apps rather than specific apps, but the performance of feature extraction varied considerably from app to app (Guzman & Maalej, 2014).

Here, we describe a competitive analysis using topic modeling and sentiment analysis on user reviews of messaging apps. Distinct from previous studies (Li et al., 2017; Shah et al., 2019; Su et al., 2019), the present work considers all of the three critical issues. Based on this concept, this work set out to investigate the usefulness of topic modeling and sentiment analysis in terms of uncovering competitive insights from user reviews of messaging apps. Three research questions need to be addressed for achieving this objective: (1) the performance of topic modeling, (2) the performance of sentiment analysis, and (3) the competitive intelligence based on topic modeling and sentiment analysis. This work is expected to provide a possible solution for existing players and new entrants of messaging apps to obtain competitive intelligence from user reviews of peer apps.

In the present work, based on text-mining user reviews of four iOS messaging apps, we adopted a topic model to extract the shared app aspect-related topics across the four competitive apps, employed a lexicon-based tool for sentiment analysis to adjust the sentiment score for each review, and performed a competitive analysis using the extracted topics and the sentiment scores. The results show that the performance of topic modeling and sentiment analysis is promising, and that the competitive analysis reveals meaningful insights into several facets.

Grounded on the objective of our research, the remainder of this work is organized as follows. The next section reviews previous studies regarding competitive analysis by text-mining UGC, topic modeling and

sentiment analysis. The methodology section introduces our research framework and describes the process of domain-specific text preprocessing, topic modeling, sentiment analysis and competitive analysis. Thereafter, we present the results and discuss our findings in section 4. Subsequently, section 5 concludes our research. Finally, we discuss the limitations of our research and provide recommendations for future works in the last section.

## 2 RELATED WORKS

We started by reading research papers in reference to competitive analysis by text-mining UGC. These studies took us to another two research fields, topic modeling and sentiment analysis. Research materials in these two domains were then reviewed, and they indeed provided a significant direction for the construction of our methodological framework.

### 2.1 COMPETITIVE ANALYSIS BY TEXT-MINING UGC

Hotel industry has been leveraging the feedback and content generated by customers to gain competitive intelligence. [Gémar and Jiménez-Quintero \(2015\)](#) text-mined what people discussed about a total number of 83 hotel brands on social media such as Twitter, Facebook, LinkedIn and YouTube from three dimensions, i.e., sentiments, passion and reach, and examined the correlation between these dimensions and the return on equity using sample data consisting of hotel information and financial statistics. Their results indicate that hotels can improve their financial performance by analyzing the content on social media. Moreover, based on a framework of strengths, weaknesses, opportunities and threats, [Amadio and Procaccino \(2016\)](#) employed text mining and visualization tools to analyze TripAdvisor reviews of three hotels located in New York City. Their study suggests that such analysis can reveal previously unknown but valuable competitive insights, which assist hoteliers in taking appropriate strategic and operational actions. [Hu and Trivedi \(2020\)](#) also analyzed online reviews on TripAdvisor. Differently, they implemented content analysis and repertory grid analysis to explore the detection of brand performance, competitive landscaping and development of competitive strategies by text-mining customer reviews of six international hotel brands. In their study, customer preferences towards hotel attributes were used to detect brand positioning and competitive groups, and customer expectations and perceptions contributed to the development of competitive strategies for identified competitive groups.

Besides hotel industry, restaurant industry and retail business are also able to obtain competitive intelligence from customer feedback. [He et al. \(2013\)](#) quantitatively analyzed the number of followers, postings, comments, shares and likes on Facebook and Twitter sites of three major pizza chains and text-mined the wall posts on their Facebook pages and the tweets on their Twitter sites. Different patterns on Facebook pages and on Twitter sites were found for the three pizza chains. Their results affirm the value of social media data in competitive analysis and indicate that social media of these businesses positively promote the customer engagement. Moreover, [Kim and Jeong \(2018\)](#) conducted a competitive analysis for two ramen rivals in Korean market by combining sentiment analysis of the online UGC on blogs and forums of the two ramen brands as well as statistical analysis to examine the correlation between market share gap and UGC volume, time-series sentiment comparison and customer sentiments towards product features. Significant gaps were shown in UGC volume and customer sentiments of the two ramen brands, clearly indicating that one is a market leader and the other is a market follower. Furthermore, the comparative relations extracted from online customer reviews of competing products ([Xu et al., 2011](#)) and restaurants ([Gao et al., 2018](#)) also reveal competitive insights. Using a graphical model based on the two-level conditional random fields ([Lafferty et al., 2001](#)), [Xu et al. \(2011\)](#) extracted comparative relations from a corpus of Amazon customer reviews and built comparative relation maps for visualization. Their

approach was capable of identifying potential operation risks, which could further guide product and marketing strategies. [Gao et al. \(2018\)](#) presented a model to mine aspect-oriented comparative relations from online reviews of restaurants and used the comparative relations to create three types of comparison relation networks, which respectively help restaurants to understand their market positioning, identify top competitor and recognize competitive strengths and weaknesses. These insights were expected to help the restaurants to develop a better service improvement strategy. Rather than using comparative opinions, [Wang et al. \(2018\)](#) performed a topic analysis by applying the latent Dirichlet allocation (LDA) ([Blei et al., 2003](#)) to online reviews of two wireless mice and two oil diffusers sold on Amazon, and the product strengths and weaknesses were effectively identified in their analysis. Different from the LDA as an unsupervised method, supervised learning was implemented for identifying competitors from the UGC on social media sites of automotive products ([Liu et al., 2019](#)). In addition to a competitor identifier, their approach also contains a domain-specific sentiment lexicon for measuring customer attitude. Competitive advantages and disadvantages were revealed based on the detected competitors and their customer attitude.

Mobile apps can also benefit from user reviews to gain competitive insights. [Li et al. \(2017\)](#) presented an approach to compare apps by identifying review sentences with comparative opinions and matching app alias in reviews. Their experiments on five million user reviews from Google Play suggest the effectiveness of the proposed method in terms of identifying meaningful comparisons across mobile apps. Similarly, CompetitiveBike is a system to predict the popularity of bike-sharing apps by analyzing data from various sources, which include microblog posts with strong comparative opinions ([Ouyang et al., 2019](#)). These approaches ignore numerous reviews and posts that are not expressed in a strongly comparative manner. A possible solution is the tool developed by [Shah et al. \(2019\)](#), which automatically classifies user reviews into five types, extracts app features using the rule-based Simple Approach for Feature Extraction (SAFE) ([Johann et al., 2017](#)) and generates summary for developers to view the feature-related information of their own app as well as competing apps. However, their results show that the SAFE suffered from low precisions in app feature extraction from user reviews, and thus manual filter is necessary to improve the performance. Furthermore, [Su et al. \(2019\)](#) combined the LDA model and sentiment analysis to analyze user reviews of similar apps and matched the extracted topics across the similar apps based on their semantic similarities. Their results are encouraging in topic coherence and sentiment analysis.

Within this context, the present work expands [Su et al.'s \(2019\)](#) study by applying topic modeling to discover the shared topics from user reviews of messaging apps without topic matching. We also conducted a competitive analysis based on the extracted topics and sentiment analysis, which was not further investigated in their study.

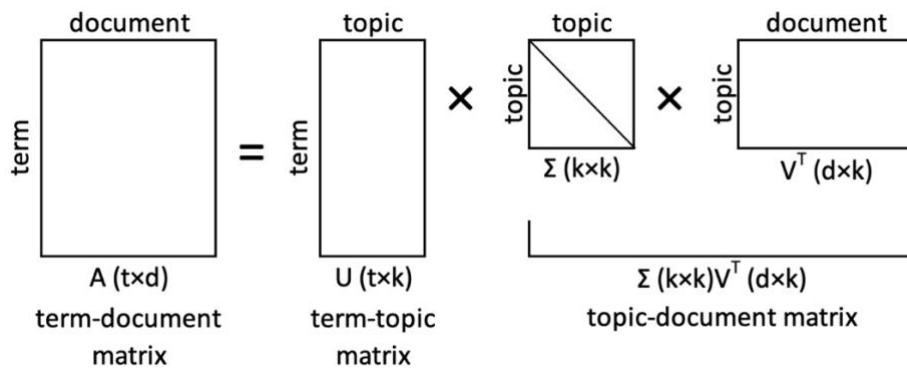
## 2.2 TOPIC MODELING

Topic modeling is a statistical tool for discovering the latent topics from many documents of texts without any prior annotations or labeling ([Blei, 2012](#)). This technique has been widely adopted in natural language processing concentrated on semantic analysis and text mining for analyzing social media content ([Hong & Davison, 2010](#)) and bioinformatics data ([Liu et al., 2016](#)).

The latent semantic analysis or indexing (LSA/LSI) (Deerwester et al., 1990) is an early topic model proposed for automatic information indexing and retrieving based on singular value decomposition. Most traditional information retrieval approaches rely on a lexical match between query words and those in queried documents, but different words can express the same concept and most terms have varying meanings, thereby returning a large amount of irrelevant information during the query (Deerwester et al., 1990; Dumais, 2005). In LSA, terms and documents are mapped to a semantic space wherein closely related terms and documents are positioned near one another, and the singular value decomposition arranges this semantic space in a way that major associative patterns are extracted from the data, and unimportant noises are ignored. As a result, documents with no co-occurred words may still end up having similar contexts. To extract  $k$  topics from a set of documents of texts, LSA decomposes the term-document matrix  $A_{t \times d}$  into a multiplication of three matrices (Figure 1). The term-document matrix is the matrix of weight or frequency representation of the text-based data, and each column in the term-topic matrix  $U_{t \times k}$  represents a topic. The topic-document matrix is structured by the product of the diagonal matrix of singular values  $\Sigma_{k \times k}$  and the transpose of the topic-document matrix  $V^T_{d \times k}$ .

**Figure 1**

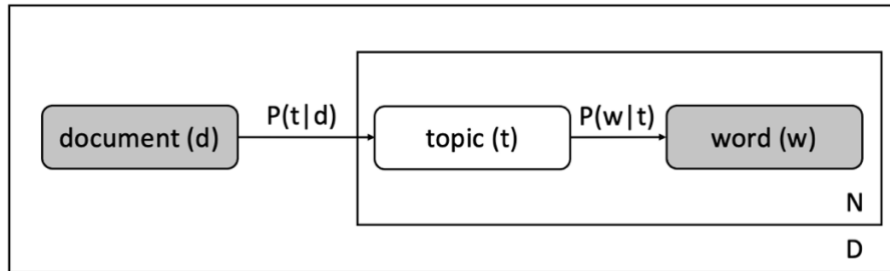
*The singular value decomposition of LSA with  $k$  topics*



Hofmann (1999) argued that LSA does not have a sound foundation in statistics and introduced the probabilistic LSA/LSI (pLSA/pLSI), also known as the aspect model. The aspect model is a statistical method based on the concept that a topic distribution  $P(t|d)$  exists in each document  $d$  of a total number of  $D$  documents, and in a total number of  $N$  words in each document  $d$ , each topic is determined by a word distribution  $P(w|t)$  (Figure 2). Both topic and word distributions follow the multinomial distribution. This approach of topic modeling changed from dimensionality reduction methods to probabilistic modeling.

**Figure 2**

*Illustration of the aspect model*

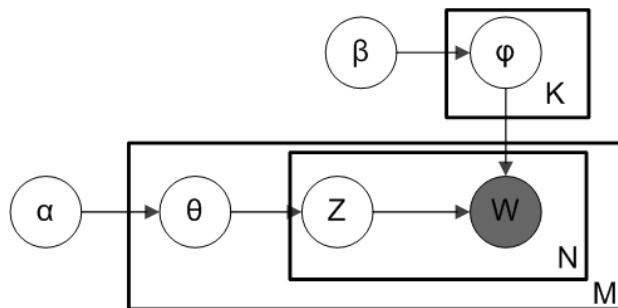


*Note.* Adapted from Probabilistic latent semantic analysis — *PLSA* by Serg Karpovich, 2013, Wikimedia Commons ([https://commons.wikimedia.org/wiki/File:Вероятностный\\_латентно-семантический\\_анализ.png](https://commons.wikimedia.org/wiki/File:Вероятностный_латентно-семантический_анализ.png)). CC BY-SA 3.0.

Unfortunately, pLSA is not capable of assigning probability to previously unseen documents and is prone to overfitting due to the growth in parameters with the increasing number of documents and words (Blei et al., 2003). To address these problems, Blei et al. (2003) presented the latent Dirichlet allocation (LDA), which is a hierarchical Bayesian model of three levels, namely document, topic and word. According to their study, a document contains multiple topics with different probabilities  $\theta$ , whose distribution follows the Dirichlet distribution, and another Dirichlet distribution also applies in the probability distribution of words  $\varphi$  in a topic (Figure 3). The parameters of the prior distributions, i.e., Dirichlet distributions, of topic distribution  $\theta$  and word distribution  $\varphi$  are  $\alpha$  and  $\beta$  respectively. Compared with the pLSA approach, the LDA method generalizes easily to unseen texts because it obtains the posterior distribution of the topic mixture weights by combining their prior distribution with the sample data rather than treats these weights as a large set of individual parameters (Blei et al., 2003).

**Figure 3**

*Illustration of the latent Dirichlet allocation*



*Note.* From *Plate notation of the Smoothed LDA Model* by Slxu.public, 2009, Wikimedia Commons ([https://commons.wikimedia.org/wiki/File:Smoothed\\_LDA.png](https://commons.wikimedia.org/wiki/File:Smoothed_LDA.png)). CC BY-SA 3.0.

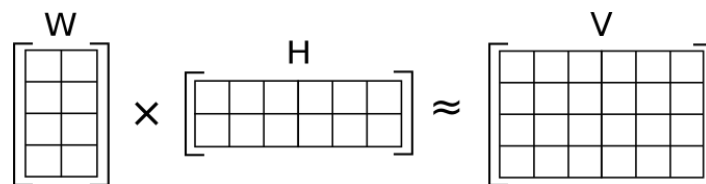


Some derivations based on the LDA method were presented for specific tasks and improvements. Combining the traditional LDA and a statistical process, [Blei et al. \(2010\)](#) created the Hierarchical latent Dirichlet allocation (hLDA) to learn topic hierarchies from complex data. In hLDA, the latent topics are structured in a tree where each node represents a topic with its topic terms. [Agarwal and Chen \(2010\)](#) proposed the matrix factorization through LDA (fLDA) to predict user ratings in recommender systems such as content recommendation, ad targeting and web search, whose items are articles, ads and web pages respectively. The fLDA method regularizes the item factors through LDA priors. In the domain of mobile apps, [Park et al. \(2015\)](#) presented the AppLDA topic model that retrieves shared topics from app descriptions and user reviews and discards reviews which only contains misaligned topics with app descriptions. This topic model aims to find out the key aspects of apps and interrelate the vocabulary between app developers and users. Moreover, to handle the restriction due to a single Dirichlet prior distribution for topic proportions, [Chien et al. \(2018\)](#) introduced the latent Dirichlet mixture model (LDMM), which allocates multiple Dirichlet prior distributions to learn the latent topics and their proportions. Besides unsupervised tasks such as topic modeling and document clustering, the LDMM was also extended to a supervised LDMM for document classification ([Chien et al., 2018](#)).

Another common topic model is the non-negative matrix factorization (NMF) ([Lee & Seung, 1999](#)), whose development is relatively independent of the aforementioned topic models. This model was initially introduced for learning parts of faces and semantic features of text. The basic idea of NMF is finding two non-negative matrices whose product approximates a given non-negative matrix, thereby factorizing the original non-negative matrix  $V$  into a basis matrix  $W$  and a coefficient matrix  $H$ , such that  $V \approx WH$  ([Figure 4](#)). The column vector of the original matrix  $V$  is the weighted sum of all the column vectors in the left basis matrix  $W$ , and the weight coefficient is the element of the corresponding column vector in the right coefficient matrix  $H$ . Both LSA and NMF share the key idea of matrix factorization and dimensionality reduction. However, singular vectors decomposed from LSA can contain negative values, while NMF has non-negative constraints and its positive and zero coefficients are more in line with human cognitive process when it comes to interpreting the importance of the words in extracted topics ([Lee et al., 2009](#)).

**Figure 4**

*Illustration of the non-negative matrix factorization*



*Note.* From *Illustration of approximate non-negative matrix factorization (NMF)* by Qwertyus, 2013, Wikimedia Commons (<https://commons.wikimedia.org/wiki/File:NMF.png>). CC BY-SA 3.0.

The LDA method is currently considered to be one of the most popular topic models and has been widely used for extracting useful information from app-related posts and reviews (e.g., [Iacob & Harrison,](#)

2013; Guzman & Maalej, 2014; Ouyang et al., 2019; Su et al., 2019). However, NMF far outperformed LDA in terms of execution time, while the accuracy difference between these two models was trivial (Truică et al., 2016). Although the NMF approach preceded the introduction of LDA, some studies have shown its effectiveness in extracting topics from an informal textual content. O’Callaghan et al. (2015) analyzed topics extracted by NMF-based methods and LDA-based methods in six corpora with a total number of 501,743 textual documents and found that topics produced by NMF regularly have higher coherence than those generated by LDA, especially for niche or non-mainstream corpora. This result is in line with Contreras-Piña and Ríos’s (2016) conclusion that NMF extracted more useful and coherent topics than LSA and LDA based on an experiment on a dataset of 21,863 consumer complaints about department store’s credit cards. Moreover, according to an experiment on 57,934 user reviews of a popular e-commerce app released on Google Play, NMF obtained a better solution compared to LDA in modeling topics from app reviews (Suprayogi et al., 2018). The recent study of Albalawi et al. (2020) shows both LDA and NMF approaches delivered more meaningful topics than LSA, random projection and principal component analysis when dealing with short texts such as comments and reviews.

User reviews of messaging apps usually contain informal expressions. Based on the existing studies, we selected the NMF method as the topic model for our research comprehensively considering effectiveness and efficiency. For user reviews, the data matrix contains non-negative values when each term in the review texts is properly represented, e.g., by scores of term frequency-inverse document frequency (TF-IDF).

### 2.3 SENTIMENT ANALYSIS

Sentiment analysis refers to the computational study of human opinions, attitudes and emotions towards entities such as individuals, events and topics, and aims to discover opinions in texts, identify the emotions these opinions express and classify their sentiment polarity (Medhat, 2014). To achieve sentiment classification, both supervised learning methods and unsupervised approaches can be applied (Liu, 2012). Supervised learning methods require training data with sentiment labels. In real practice, data sources from user-generated posts and reviews do not include such labels. Classifying the sentiment polarity or strength scale to UGC has become a popular research topic in recent years. According to Liu (2012), sentiment words dominate the sentiment polarity, and thus these sentiment words and phrases may be utilized for sentiment classification in an unsupervised manner.

One of the unsupervised approaches to sentiment analysis for UGC is the lexicon-based method, which requires a predefined lexicon. The General Inquirer (Stone et al., 1966) is a lexical set with syntactic, semantic and pragmatic information of part-of-speech tagged words and gives labels of the sentiment polarity to most of its included words. The Multi-Perspective Question Answering (MPQA) subjectivity lexicon (Wilson et al., 2005) also provides the same structural information as the General Inquirer. Moreover, Wilson et al. (2005) included the additional subjectivity level of a word or a phrase with a label for strong or weak. Bradley and Lang (1999) created the Affective Norms for English Words (ANEW), a dictionary in which 1,034 English words are rated in terms of valence, arousal and dominance on a continuous scale between 1 and 9. The three rated dimensions are based on Osgood et al.’s (1957) theory of emotions. In their theory, valence (or pleasantness) and arousal (the intensity of excitement)

are the principal dimensions, while dominance (or control) is a less strongly-related dimension when it comes to emotions invoked by a word. Similar to ANEW (Bradley & Lang, 1999), SentiWordNet (Esuli & Sebastiani, 2006; Baccianella et al., 2010) also assigns three sentiment scores to each synonym set (synset), albeit from three different aspects regarding positivity, negativity and neutrality. This lexical resource is constructed on WordNet (Miller, 1995; Fellbaum, 1998) and publicly available in the Natural Language Toolkit (NLTK) (Bird et al., 2009) for research purposes. Furthermore, Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2003) provides a proprietary dictionary that organizes each word or word stem across one or more psychologically relevant categories, which can infer positive or negative emotions. With the rapid development of social media, more and more cyber related words appear in microblogs. Nielsen (2011) stated that some of the aforementioned General Inquirer, MPQA subjectivity lexicon, ANEW and SentiWordNet do not incorporate strong obscene words and Internet slangs, and thus he presented a new ANEW, a Twitter-based sentiment word list including cyberslangs and obscene words. In his comparative experiment, the new ANEW exceeded the General Inquirer, MPQA subjectivity lexicon and ANEW, but all these lexicons did not perform as well as SentiStrength (Thelwall, 2013), a lexicon-based tool for sentiment analysis.

According to Thelwall (2013), SentiStrength is a sentiment analysis tool for classification of social web texts. This tool uses words and word stems from the existing LIWC and General Inquirer as well as special words and phrases widely used on social media, and sentiment scores of these words are annotated by humans and improved with machine learning methods. Also, some rules were constructed to cope with non-standard textual expressions particularly in social media such as emoticons, emphasized punctuations and intended misspellings. The tested cases show that SentiStrength performed well on a wide range of social media texts, but less well on texts with ironic and sarcastic expressions (Thelwall, 2013). SentiStrength has been used for aspect-based sentiment analysis for user reviews of mobile apps (Guzman & Maalej, 2014), for ranking product aspects from online reviews (Wang et al., 2016) and for sentiment measurement for user comments on social media (He et al., 2016).

Similar to SentiStrength, the Valence Aware Dictionary for sEntiment Reasoning (VADER) (Hutto & Gilbert, 2014) is another lexicon-based sentiment analysis tool designed for social media contexts. The human-validated sentiment lexicon of VADER exploits the existing lexicons such as the General Inquirer, LIWC and ANEW, incorporates additional terms commonly used in microblogs and is built on five grammatical and syntactical rules such as punctuations, capitalization and degree modifiers. In their experiments, VADER achieved notable success in social media domain and generally outperformed a majority of the well-regarded sentiment analysis tools including the General Inquirer, ANEW, LIWC, Hu-Liu04 opinion lexicon (Hu & Liu, 2004), Word-Sense Disambiguation (Akkaya et al., 2009), SentiWordNet and SenticNet (Cambria et al., 2012). Recent studies have employed VADER for analyzing the sentiments of user reviews of mobile apps (Huebner et al., 2018; Su et al., 2019) and microblogs on Twitter (Elbagir & Yang, 2019) with encouraging performance.

SentiStrength and VADER differ in the output of sentiment scores. For each input, SentiStrength reports two independent scores of positive and negative scales based on the psychological theory of humans'

mixed emotions (Norman et al., 2011) (Thelwall, 2013), while VADER provides a compound score computed from the positive, neutral and negative scores (Hutto & Gilbert, 2014).

Another earlier introduced dictionary-based tool for extracting sentiment from texts is the Semantic Orientation CALculator (SO-CAL) (Taboada et al., 2011). Different from word-based approaches using adjectives (e.g., Whitelaw et al., 2005) or adjectives and adverbs (e.g., Benamara et al., 2007) to infer the emotional orientation, SO-CAL exploits words including adjectives, verbs, nouns, and adverbs to calculate the sentiment polarity and strength. Apart from this extension of parts of speech (POS), SO-CAL also incorporates a dictionary of intensifiers and a refined negation approach. Their results show that SO-CAL achieved consistent performance on completely unseen texts across domains, different from SentiStrength and VADER designed for social web texts.

User reviews are one type of UGC, whose language expressions are similar to user posts and comments on social media. In the present work, we considered sentiment analysis tool with a focus on social media texts and included the recently introduced VADER to be part of the sentiment analysis of our research since the VADER compound scoring method is more appropriate to our further calculation of the final sentiment scores.

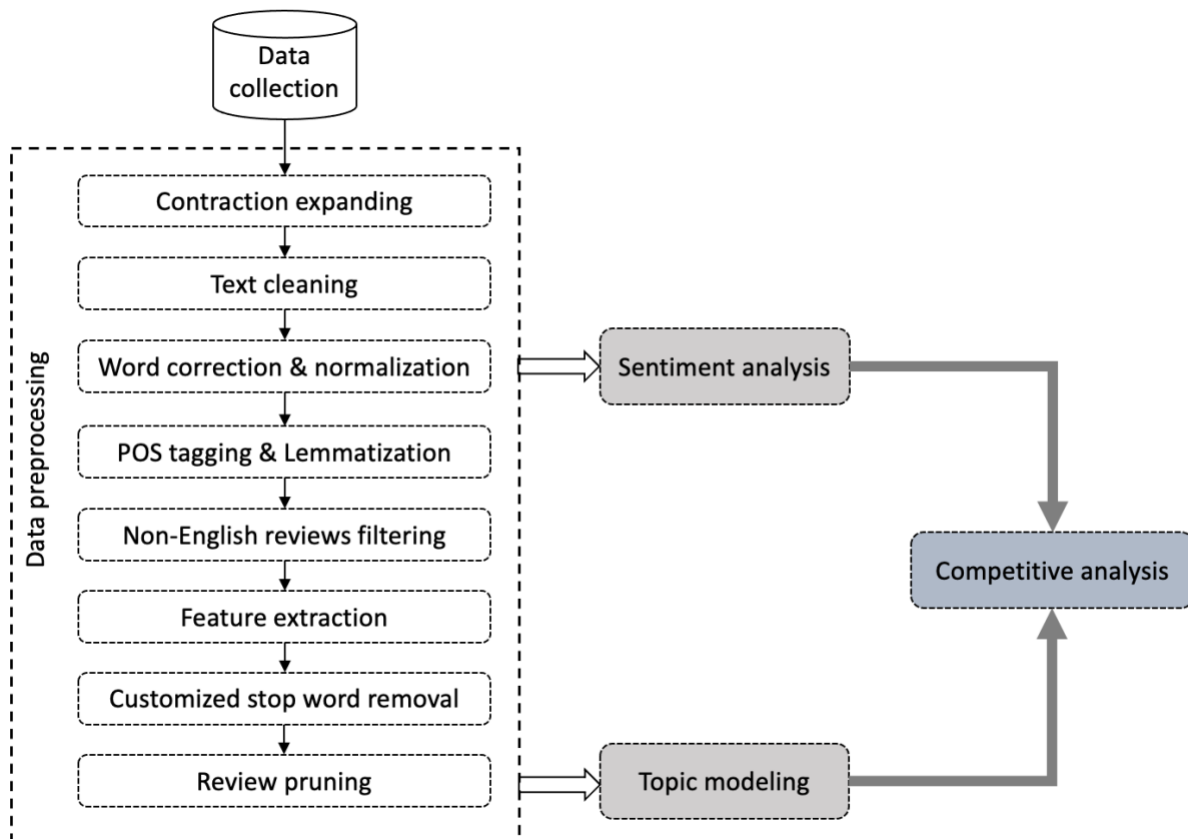
### 3 METHODOLOGY

This section introduces the generic architecture of our research focusing on three key research issues: topic modeling, sentiment analysis and competitive analysis (Figure 5).

As depicted in Figure 5, our framework starts with data collection, followed by a series of domain-specific data preprocessing steps including contraction expanding, text cleaning, spelling normalization, POS tagging, word lemmatization, non-English reviews removal, app feature extraction, removal of customized stop words and review pruning. After this, a topic model takes the extracted feature terms as input to find underlying app aspect-based topics. In parallel, partially preprocessed review sentences were used to compute their sentiment scores. In the end, we summarized the extracted topics and the sentiment scores to conduct a multi-faceted competitive analysis. In addition to statistical summaries, we also created comparative visualizations to better reveal competitive insights into several aspects.

**Figure 5**

*The research framework of competitive analysis using topic modeling and sentiment analysis*



### 3.1 DATA COLLECTION

The app review management platform AppFollow ([Appfollow, n.d.](#)) aggregates user reviews from iOS, Android, Microsoft and Amazon app stores. We selected the United States as the country and separately exported user reviews of four iOS mobile apps, namely Messenger, WhatsApp Messenger, Signal - Private Messenger and Telegram Messenger, between June 1, 2020 and May 31, 2021. Since the app names of WhatsApp Messenger, Signal - Private Messenger and Telegram Messenger carry the word “Messenger”, which is also the name of the other app Messenger, we replaced the three app names with “WhatsApp”, “Signal” and “Telegram” respectively to clarify the app names and minimize redundancy.

A total number of 27,479 reviews were collected. [Table 1](#) lists the number of reviews and review percentage for each app.

**Table 1**

*Information of app and user reviews*

App Name	App Company	# of Reviews (%)
Messenger	Facebook, Inc.	12,346 (44.93%)
WhatsApp	WhatsApp Inc.	10,513 (38.36%)
Signal	Signal Messenger, LLC	2,633 (9.58%)
Telegram	Telegram FZ-LLC	1,987 (7.23)
TOTAL		27,479 (100%)

The exported datasets were concatenated into a single dataset with 26 fields. The fields regarding the review datetime, app name, user rating and review content were selected for further analysis. [Table 2](#) provides a sample of the dataset with selected fields.

**Table 2**

*Sample of dataset with selected fields*

Date	AppName	Rating	Review
2020-06-01 01:10:35	Messenger	1	Can't send pictures or videos
2020-06-01 00:06:08	WhatsApp	5	Still awesome
2020-06-01 01:17:05	Signal	4	i love the app since i realized i needed the privacy from hangouts and zoom, but whenever i try to add someone to a groupchat, it shows up as “error user already in group” when they're not? is anyone else having this problem?
2020-06-01 04:14:59	Telegram	5	The best app I've ever seen

## 3.2 DATA PREPROCESSING

### 3.2.1 Contraction expanding

Contractions are widely used in informal writing, and user reviews are no exception. The two terms “doesn't” and “does not” have exactly the same meaning, but they are two completely different tokens for a computer. Such contractions increase the dimensionality of the document-term matrix for further topic modeling. To reduce the redundancy in the data, we expanded the common contractions and slangs using the Contractions Python library ([Kooten, n.d.](#)) such that, e.g., “doesn't” was expanded to “does not”, “could've” to “could have”, and “wanna” to “want to”. This step also prevented the subsequent text cleaning from generating many wrong spellings such as “doesn t” and “could ve” after removing the apostrophes. In addition, this Python library is also capable of correcting common typos related to contractions such as “didnt” and “cant”. This kind of typos would be automatically expanded.

### 3.2.2 Text cleaning

User reviews are usually mixed-cased and comprise many non-textual characters such as punctuations and digits as well as characters in foreign languages. These characters provide least useful information about the app features and user attitude but introduce many noises to the topic model. As a result, we lowercased all review texts and then removed all URL's, newline characters, punctuations, emojis, digits and non-English characters.

Also, reviews with only one word such as “awesome” and “ugh” were dropped. These short reviews are usually praise, critiques or modal particles without mentioning any specific app features, and thus convey meaningless information for competitive analysis. A total of 1,952 such reviews was removed and 25,527 reviews remained at the end of this step.

### 3.2.3 Word correction and normalization

User reviews tend to involve spelling errors, variant forms of English spellings and informal abbreviations. [Gu and Kim's \(2015\)](#) typo list only collects the common typos in user reviews of general apps, and most of their typos could be corrected by the Contractions Python library. Moreover, their typo list does not incorporate different forms of English spellings, which express the same meaning but represent entirely different things for a computer just like contractions. Since we were dealing with a specific type of apps, we performed a domain-specific correction and normalization of words to reduce the noises for further POS tagging, topic modeling and sentiment analysis.

This sub-section includes checking the spelling property of each unique word, normalizing British spellings to American spellings, and correcting English misspellings and abbreviations.

#### 3.2.3.1 Spell check

Before performing the spelling normalization and correction, we needed to identify the spelling properties of each unique word. The PyEnchant ([Kelly, 2011](#)) spellchecker was adopted for this task. This Python library is capable of recognizing different varieties of the English language such as American

English, British English and Canadian English. The same word in different varieties is understandable for humans but noisy for computers. The two common American and British varieties were considered in our spell check.

We computed the frequency of each unique word in the corpus and sorted this frequency in descending order. The unique uppercase words were then passed to PyEnchant for checking if each of them is a correctly spelled English word, in both American and British spelling. We used capital words because all words were lowercased during the text cleaning, and PyEnchant is case-sensitive. For instance, “english” is not a correct spelling, but “English” or “ENGLISH” is. Finally, for each unique word, the PyEnchant spellchecker reported “True” or “False” in both American and British spellings.

### 3.2.3.2 British-American spelling normalization

For words in *True* British spelling but *False* American spelling, we normalized them to American spellings based on the British spelling dictionary of an American-British English translator ([Hyperreality, n.d.](#)), e.g., “behaviour” to “behavior”. Some of the *True* British spellings were not found in this dictionary, so we manually added the American spellings ([Table 3](#)). Finally, we replaced all the British spellings in the cleaned review texts with their American counterparts.

**Table 3**

*Additional British-American spellings*

British spelling	American spelling
amongst	among
learnt	learned
customisation	customization
acknowledgement	acknowledgment

### 3.2.3.3 Spelling correction and abbreviation expansion

Words with both *False* spellings are usually misspelled English words, informal English abbreviations or non-English words. In this step, we only corrected the English misspellings and informal English abbreviations, e.g., “recieve” to “receive” and “pls” to “please”. The detailed steps were selecting the both *False* words, going through those with a frequency greater than two and manually assigning the correct form to each word if the word obviously resembles an English word or abbreviation ([Table 4](#)).



**Table 4***Manual spelling corrections*

Word	Manual correction
fb	facebook
pls, plz, pleasee, pleaseee, plzzz, pleaseeee	please
idk	receive do not know
ui	user interface
messenger, messanger, massenger, messnger, mesenger, mesengger	messenger
ppl, ppls	people
msgs, messeges, messenges	messages
cuz, coz, bcz, cus	because
receive	a lot
devs	developers
watsapp, whatsup, whatapp, whatsap, sapp, whatsapp, whatsaap, whatsapp, whatsapps, watsapp, whatsapp, whatup, whastapp, watsap, whatsapp	whatsapp
thx	thanks
mins	minutes
soooo, sooo, soo, sooooo, soooooo	so
ux	user experience
acc	account
appstore	app store
useable	usable
dev	developer
untill	until
networkmanagererror	network manager error
iam	receive am
rlly, realy	really
receive	happened
receive	receive
goin	going
faceid	face id
hav, hv	have
homescreen	home screen
dosent	does not
awsome	awesome

**Table 4** (Continued)

videocalls	video calls
stoped	stopped
abt	about
eachother	each other
stil	still
noooo	no
useing	using
wether	whether
sth	something
bt	but
lastname	last name
freind	friend
thankyou	thank you
messege	message
stikers	stickers
fav	favorite
redownloaded	redownloaded
storys	stories
groupchat	group chat
usefull	useful
unistall	uninstall
lastest	latest
background	background
aswell	as well
andriod	android
jus	just
gettin	getting
freinds	friends
allways	always
infos	information
looong	long
befor	before

---

We also noticed that some words with both *True* spellings are widely-used abbreviations in modern life. Such common abbreviations were manually normalized to their expanded form (Table 5).

**Table 5**

*Abbreviation expansions*

Abbreviation	Expansion
pic, pics	picture
info	information
ad, ads	advertisement
msg	message
pc	computer
bc, bcs	because
ap, app, apps	application
sec, secs	second
min	minute
hr, hrs	hour
yr, yrs	year

In the end, we matched all these misspellings and abbreviations (Table 4, Table 5) in the review texts and substituted them with their manual corrections or expansions.

### 3.2.4 POS tagging and lemmatization

POS tagging is a widely adopted technique for labeling the property of each word in a sentence. For instance, POS tags for sentence “This application is amazing” are DT-NN-VBZ-JJ in sequence. The “VBZ” tag indicates a verb of third-person singular in present tense. In general, nouns, verbs and adjectives contribute significantly to the expression of app features and user attitude, while words like pronouns or prepositions usually do not carry meaningful value in terms of user opinions. With the POS tag for each word, we can extract meaningful words based on specific POS tags.

Lemmatization is a process to resolve each word in texts to its canonical form, which is also called a lemma. For example, the sentence “The buttons are not working smoothly” can be lemmatized to “the button be not work smoothly”. Since “buttons” and “button” suggest a very close semantic context, the lemmatization process also contributes to the dimensionality reduction of the document-term matrix for further topic modeling the same way contraction expansion as well as spelling correction and normalization do.

We employed Stanford Core NLP (Manning et al., 2014) to generate the POS tag and lemma in the form of *{lemma, POS tag}* for each word in the review texts. For instance, the word “messages” would be

transformed to  $\{message, NNS\}$ , where *message* is the base form of *messages* and *NNS* means a plural noun.

### 3.2.5 Non-English reviews filtering

The step of text cleaning only removed non-English characters. Non-English reviews could contain all words made up of English letters. These user reviews are not the research object in the present work and are likely to negatively impact the performance of further topic modeling and sentiment analysis. Consequently, it is necessary to filter out as many non-English reviews as possible. [Truică et al. \(2015\)](#) leveraged the frequencies of stop words and diacritics in texts to automatically identify Romance languages. Their method would not work in our case for two reasons. One is that many English reviews are short texts and do not include any common stop words, e.g., the review “keep crashing”. The other reason roots in the fact that user reviews are not limited to Romance languages, and thus diacritics are not the key differentiators to separate English from other languages.

However, using the frequency of specific words in the review texts is still a feasible idea. We exploited the POS tags generated in the previous step to filter out non-English reviews. For each review, we computed the number of the POS tags “FW”, the acronym of foreign word. Reviews with a number of FW tags greater than two were dropped and those having one or two FW tags were manually checked. In the end, a total number of 1,349 non-English reviews were removed and 24,178 reviews remained.

### 3.2.6 Feature extraction

Noun and noun phrases have been used to identify the attributes of products (e.g., [Hu & Liu, 2004](#); [Archak et al., 2011](#)). However, users tend to express mobile app features using nouns (e.g., battery, screen) and verbs (e.g., freeze, crash), while adjectives and adverbs are often used to describe these nouns and verbs ([Vu et al., 2015](#)). We followed the approach adopted by [Vu et al. \(2015\)](#) to use nouns and verbs to identify app features. Lemmatized words with the POS tags in [Table 6](#) were extracted from the review texts.

**Table 6**

*POS tags for feature extraction*

POS tag	Description
NN	Noun, singular or mass
NNS	Noun, plural
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present

**Table 6** (Continued)

*Note.* Adapted from “Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision)” by B. Santorini, 1990, p. 6. ScholarlyCommons.

**3.2.7 Customized stop word removal**

Some terms in the extracted feature words provide little information about app features due to their meanings or extreme frequencies. Therefore, we created a list of stop words to minimize the noises for further topic modeling. Our list of stop words comprises five components. In addition to the default list of English stop words of NLTK and English letters that are not in the NLTK list, we included some unimportant words, domain-specific terms with extremely high frequencies in the extracted features, and indefinite pronouns, which were tagged as nouns by the POS tagger of Stanford Core NLP.

Table 7 summarizes the removal process of the customized stop words. In the first step, all stop words in the list comprising the five components were removed from the extracted feature words. In the next step, we dropped feature words with a frequency lower or equal to ten in all the extracted feature words based on Step 1. These feature words with low frequencies do not characterize app features very well.

**Table 7**

*Removal process of customized stop words*

Step	Component	Stop words
1	Generic stop words	English stop words in NLTK default list
	English letters	a-z not included in the generic stop words
	Unimportant words	minute, day, month, year, thing, wth, omg, tbh, ect, nd
	Domain-specific terms	application, facebook, messenger, whatsapp, signal,
	Indefinite pronouns	someone, nothing, everything, something, anything
2	Low frequency feature words	words with a frequency $\leq 10$

**3.2.8 Review pruning**

A total number of 2,783 reviews with extracted feature terms fewer than 2 were removed, and the remaining 21,395 reviews were kept for further topic modeling and sentiment analysis. Reviews with zero feature term such as “Still awesome” and “Very good” are uninformative for competitive intelligence at app aspect-based topic level, while a small number of reviews with a single feature term with no co-occurring feature words are mostly uninformative as well and introduce more sparsity to the topic model.

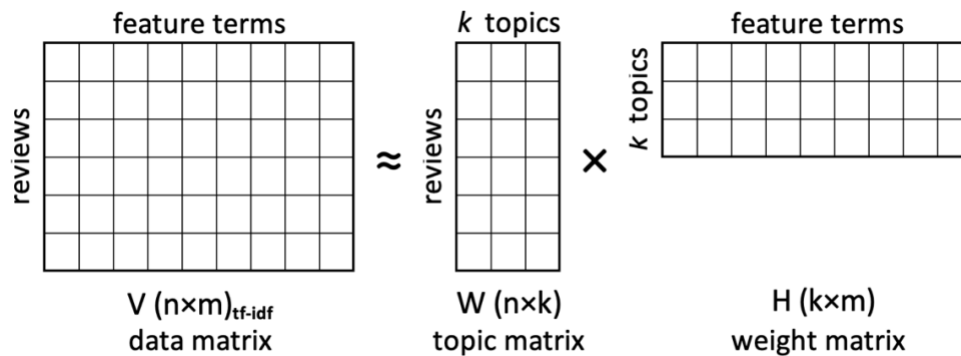
### 3.3 TOPIC MODELING

#### 3.3.1 NMF topic model

Based on the NMF form  $V \approx WH$  (Lee & Seung, 1999), Figure 6 illustrates the decomposition of topic modeling for our research. In the data matrix, each row represents a review that comprises a vector of TF-IDF scores for the extracted feature terms, i.e., nouns and verbs, while each column represents the feature term variables. The NMF topic model factorizes the data matrix into a topic matrix containing the probabilities of the  $k$  latent topics for each review and a weight matrix with the weight values of the feature terms for each topic. The number of topics to extract  $k$  is a hyperparameter to be defined.

Figure 6

*Non-negative matrix decomposition on feature terms*



We implemented the NMF topic model using Scikit-learn's (Pedregosa et al., 2011) Python library. Firstly, the feature terms were vectorized to unigram-based TF-IDF weights by Scikit-learn's `TfidfVectorizer` as the input for the topic model. Secondly, we defined  $k = 12$  as the number of latent topics to be extracted from the model. Thirdly, the Non-negative Double Singular Value Decomposition (NNDSD) method was selected for initializing the topic modeling process. This initialization method can ensure a deterministic outcome and demonstrated great efficiency when using sparse text data (Boutsidis & Gallopoulos, 2008). Lastly, we configured a maximum of five hundred iterations before the process convergence.

After processing the NMF topic model with the defined hyperparameters, we selected the top fifteen words in weight values as the topic words for each extracted topic. Thereafter, we followed Guo et al. (2017) and named each topic according to the logical relations between the selected top words and their corresponding weight values. In detail, we kept the topic labels from 0 to 11 assigned by the topic label for each topic and connected the topic label and the topic name with a hyphen "-" for each topic. Also, we separated each aspect by a vertical bar "|" if more than one aspect was mentioned in the topic.

### 3.3.2 Evaluation of extracted topics

To evaluate the performance of the NMF topic model, we randomly sampled 1% of reviews from each topic and conducted a manual analysis. We labeled “1” if a sampled review mentions any aspect within the corresponding topic name, otherwise “0”. The accuracy of the extracted topic  $t$  is calculated as:

$$Accuracy(t) = \frac{SUM(t_1)}{SUM(t)} \quad (1)$$

where  $SUM(t_1)$  expresses the count of samples with label 1 for topic  $t$  and  $SUM(t)$  is the total number of samples for topic  $t$ .

## 3.4 SENTIMENT ANALYSIS

Reviews with almost the same content might receive varying rating scores from different users. To mitigate the bias of user ratings, we included the VADER compound score to adjust the final sentiment score for each review.

### 3.4.1 VADER compound score

VADER compound score is the superlative unidimensional metric to measure the sentiment for a given sentence, and this score is calculated by summing the valence scores of each term in the VADER lexicon, weighted according to grammatical and syntactical rules, and scaled to a range between -1 (*most extreme negative*) and +1 (*most extreme positive*) (Hutto & Gilbert, 2014).

We used the Python tool developed by the authors of VADER to obtain the compound score for each cleaned review with spellings corrected and normalized. Since the user ratings follow a 1-5 rating scale, we normalized each VADER compound score  $s_{vader}$  to be between 1 and 5 to facilitate further calculation:

$$s_{vader\_normalized}(s_{vader}) = 1 + \frac{(s_{vader} - \min(S\_VADER)) * (5 - 1)}{\max(S\_VADER) - \min(S\_VADER)} \quad (2)$$

where  $\min(S\_VADER)$  and  $\max(S\_VADER)$  are respectively the smallest value and the largest value among all the VADER compound scores  $S\_VADER$ .

### 3.4.2 Weighted sentiment score

Considering both user ratings and normalized VADER compound scores, we computed the final sentiment score for each review  $r$  by using a weighted average:

$$s_{weighted}(r) = 0.5 * rating(r) + 0.5 * s_{vader\_normalized}(r) \quad (3)$$

where  $rating(r)$  corresponds to the score rated by a user for review  $r$  and  $s_{vader\_normalized}(r)$  is the normalized VADER compound score for review  $r$ .

### 3.4.3 Evaluation of sentiment analysis

Since we used the weighted sentiment scores, rather than the user ratings or normalized VADER compound scores, as the final sentiment scores for further competitive analysis, we assessed the effectiveness of the sentiment analysis from two perspectives. Firstly, we evaluated the performance of the weighted sentiment scores. Secondly, we compared the performance of the weighted sentiment scores with that of the user ratings and of the normalized VADER compound scores.

For evaluating the performance of the weighted sentiment scores, a label indicating sentiment polarity was automatically assigned to each review according to a score benchmark of 4 based on the weighted sentiment score. For example, reviews with weighted sentiment scores greater than or equal to 4 were automatically assigned a “Positive” label, otherwise “Negative”. This score benchmark was based on the assumption that reviews with a sentiment score above 4 usually express praise or friendly advice, while below 4 tend to report some issues regarding specific app features, not necessarily consisting of words linking to emotions of hate, anger or annoyance. After automatically assigning the label of sentiment polarity based on the weighted sentiment scores to each review, we randomly sampled 1% of reviews from each polarity and conducted manual labeling of “Positive” or “Negative” for each sampled review. The criterion of manual labeling conformed to the assumption for the score benchmark. An additional remark to the criterion was that reviews expressing praise first and then a shift to report issues regarding specific app aspects were manually labeled as “Negative”.

To compare the performance of the weighted sentiment scores with that of the other two scoring methods, we automatically assigned two more labels indicating sentiment polarity on the user ratings and normalized VADER compound scores to each sampled review according to the same score benchmark of 4.

At this point, each of the sampled reviews had three automatic labels of sentiment polarity on the weighted sentiment scores, user ratings and normalized VADER compound scores respectively according to the score benchmark of 4 and one manual label. We summarized the number of sampled reviews based on the labels of sentiment polarity and presented the statistical results in three confusion matrices (Table 8). Table 9 explains the terminology of TP, FP, TN and FN.

**Table 8**

*Confusion matrix for sentiment evaluation*

	Weighted sentiment scores/User ratings/normalized VADER compound scores	
	Negative (< 4)	Positive (>=4)
Negative (manual)	TN	FP
Positive (manual)	FN	TP



**Table 9***Explanation of TP, FP, TN and FN*

Acronym	Term	Description
TP	True Positive	the number of reviews with both an automatically assigned label (based on the score benchmark of 4) and a manual label to be “Positive”
FP	False Positive	the number of reviews with an automatically assigned label (based on the score benchmark of 4) “Positive” and a manual label “Negative”
TN	True Negative	the number of reviews with both an automatically assigned label (based on the score benchmark of 4) and a manual label to be “Negative”
FN	False Negative	the number of reviews with an automatically assigned label (based on the score benchmark of 4) “Negative” and a manual label “Positive”

Finally, we used the following classification metrics Precision, Recall and F1-score to respectively evaluate the performance of the weighted sentiment scores, user ratings and normalized VADER compound scores:

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \quad (6)$$

### 3.5 COMPETITIVE ANALYSIS

This section summarizes the extracted topics and the weighted sentiment scores from different perspectives for revealing competitive intelligence. The summary of review distributions and average sentiment scores indicated the overall review counts and user sentiments by topic for each app during the period between June 1, 2020 and May 31, 2021, while the sentiment evolution reflected the changes of user sentiments over time on a monthly basis.

#### 3.5.1 Review distributions and average sentiment scores

To understand the review distribution, we aggregated the number of reviews by topic for each app. The mostly discussed topics are usually the main app aspects that the users are more concerned about.

Using the weighted sentiment scores as the final sentiment scores, we also calculated the average sentiment scores by topic for each app. The average score of the topic  $t$  for the app  $a$  was calculated using the following formula:

$$Avg.SentiScore(t_a) = \frac{1}{n} \sum_{i=1}^n s\_weighted(r_{t_a})_i \quad (7)$$

where  $n$  is the total number of reviews of the topic  $t$  for the app  $a$ , and the weighted sentiment score of each review is denoted by  $s\_weighted(r_{t_a})_i$  where  $i = 1, 2, \dots, n$ .

### 3.5.1.1 Visual review distribution

To better visualize the main app aspect-based topics discussed by users, we created four pie charts which respectively show the percentages of reviews regarding each topic for the four messaging apps based on the review aggregation by topic for each app.

### 3.5.1.2 Visual comparison of average sentiment scores

We created a bar chart based on the average sentiment scores by topic for each app to visually compare the average sentiment scores of the four messaging apps at a topic level. The average sentiment scores were assumed to reflect the user satisfaction towards specific app aspect-based topics.

## 3.5.2 Sentiment evolution

We grouped the reviews by month and computed the average sentiment scores by topic for each app on a monthly basis. Moreover, to better visualize the monthly changes of user sentiments of the four messaging apps from June 2020 to May 2021, we plotted twelve line charts, each for an app aspect-based topic.

## 4 RESULTS AND DISCUSSION

After multiple steps of data preprocessing, a total number of 21,395 informative reviews remained for topic modeling and sentiment analysis, whose results are demonstrated in this section, along with the statistical and visual outcomes of competitive analysis. At the end, a discussion regarding the reasonings and interpretations behind our findings is presented.

### 4.1 RESULTS

#### 4.1.1 Topic modeling

This sub-section presents the topics extracted from the topic model and the accuracies of topic extraction in the manual evaluation.

##### 4.1.1.1 Extracted topics

Figure 7 displays the twelve topics decomposed from the NMF topic model, each with its topic label number and topic name based on the fifteen topic-words and their corresponding weight values. Most topics have one or two dominating words except for topic 2 - *(group) chat | add feature | privacy* and topic 10 - *app/account deletion | download | account/login*. These two topics hold a fair number of words with moderately decreasing importance, and thus they comprise more aspects regarding app features. The rest of the topics, though led by one or two words, are not always merely or intuitively represented by their dominating words. The other words in these topics also complete their dominating words to represent the aspect more clearly or even give information about additional app-related aspects depending on the weight values of the topic words.

Figure 7

*Topics extracted from the NMF model*

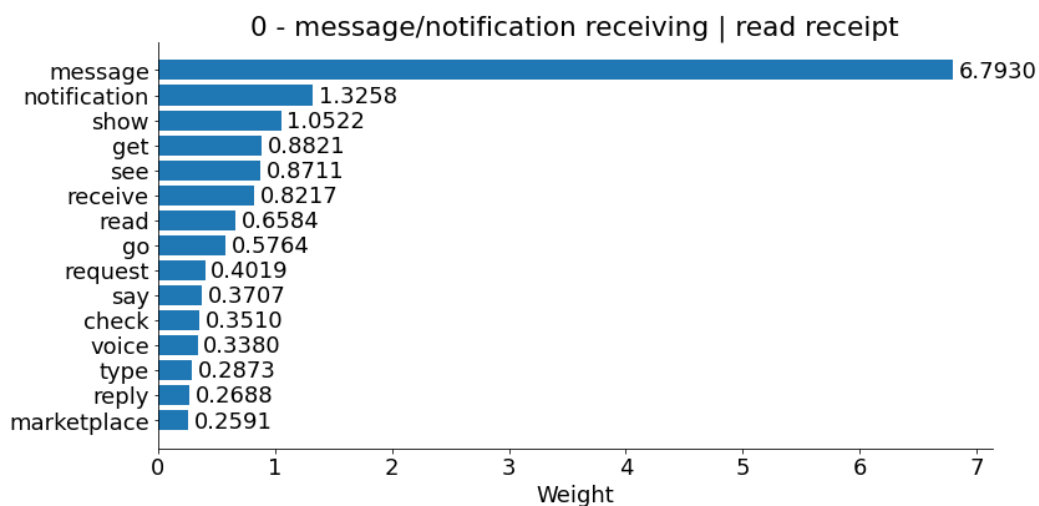


Figure 7 (Continued)

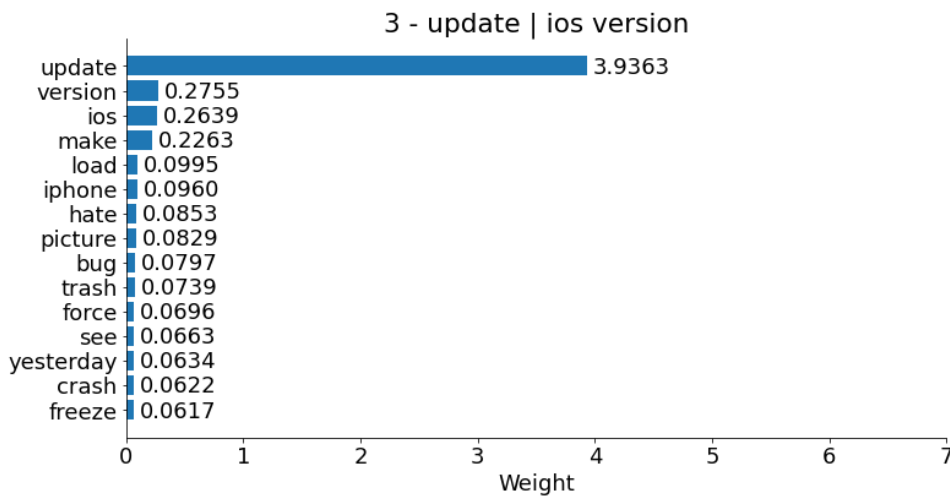
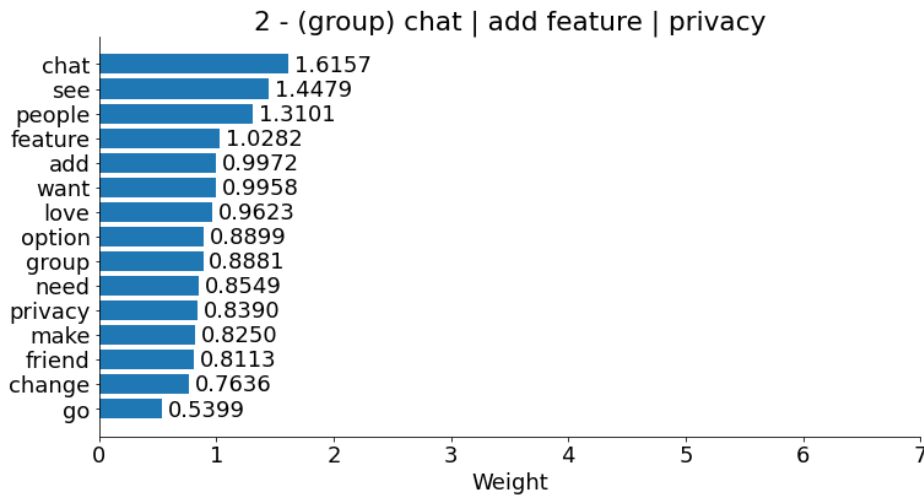
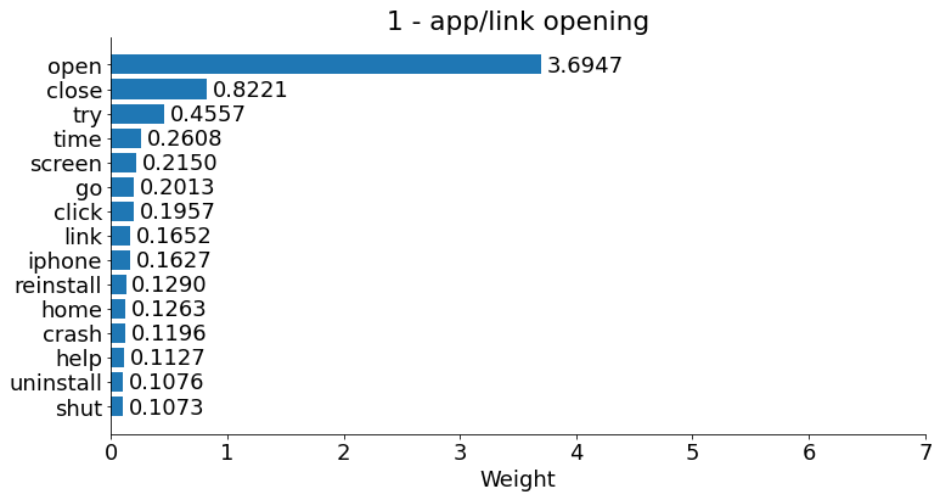


Figure 7 (Continued)

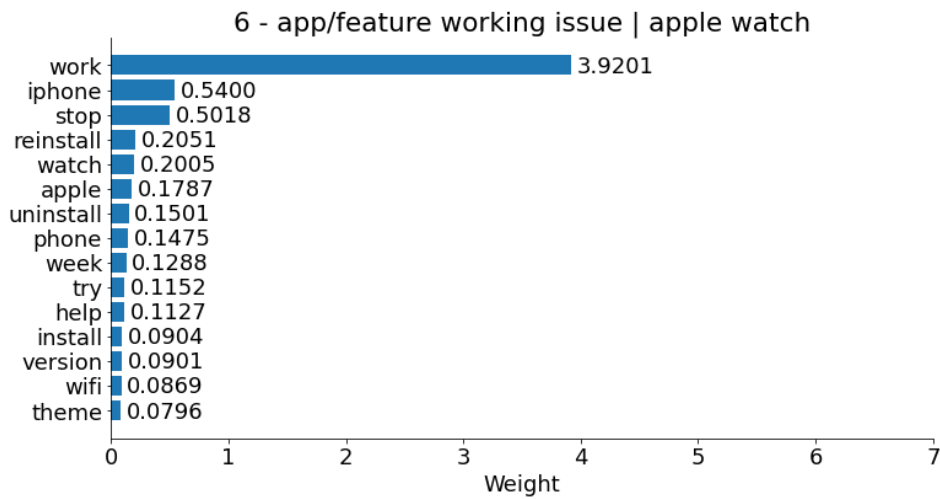
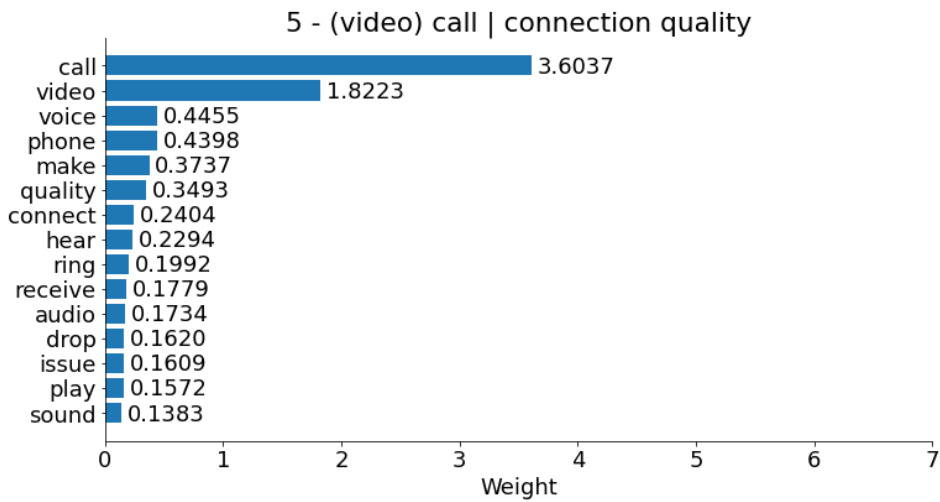
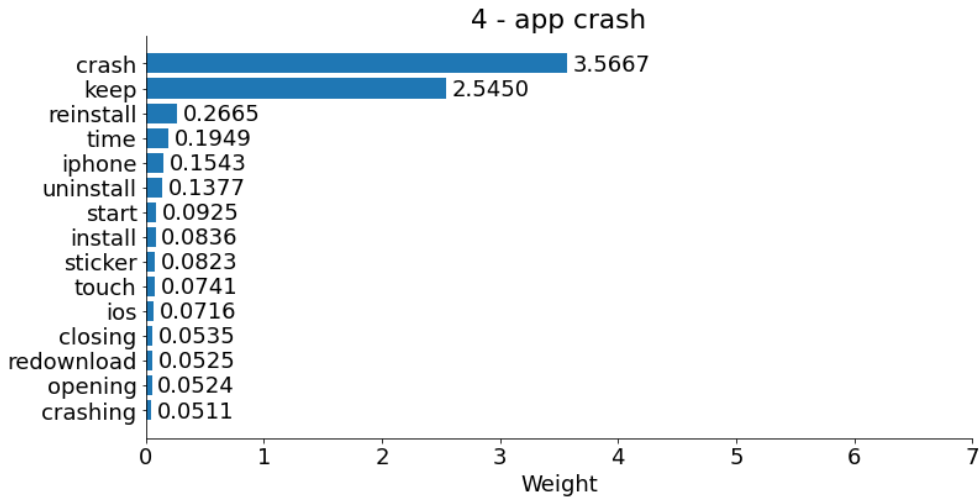


Figure 7 (Continued)

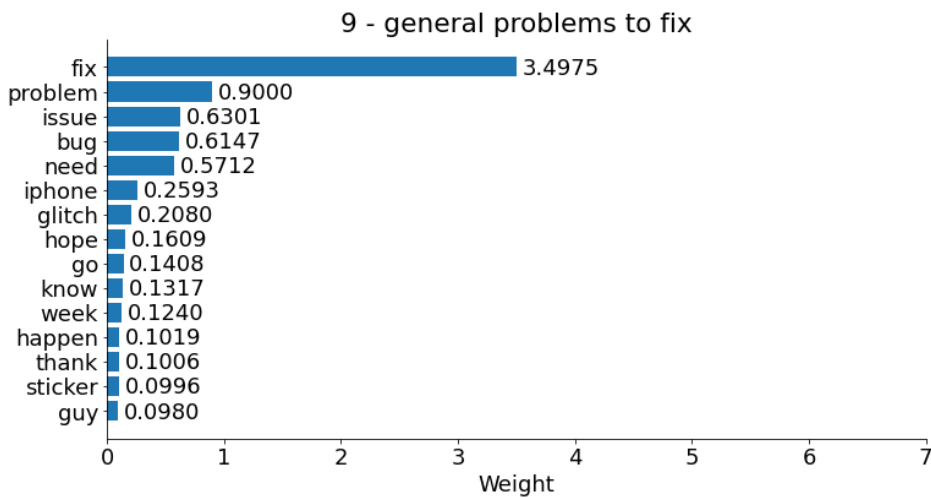
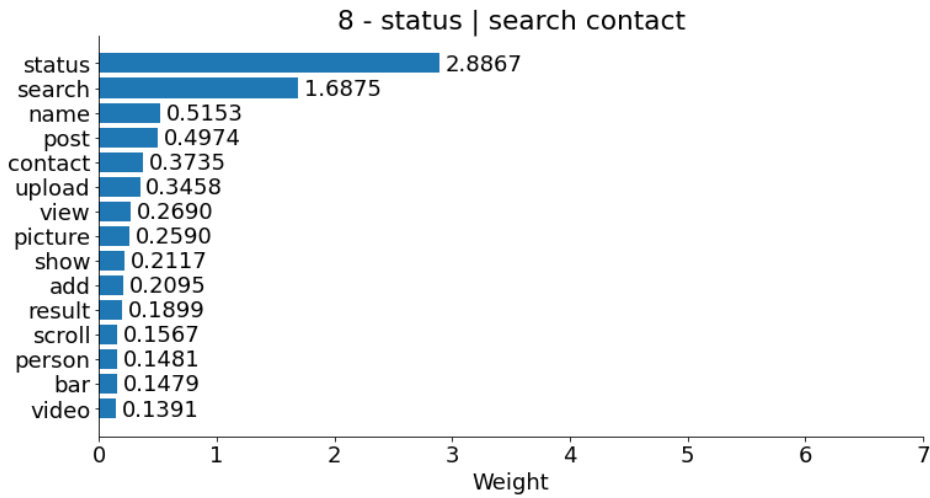
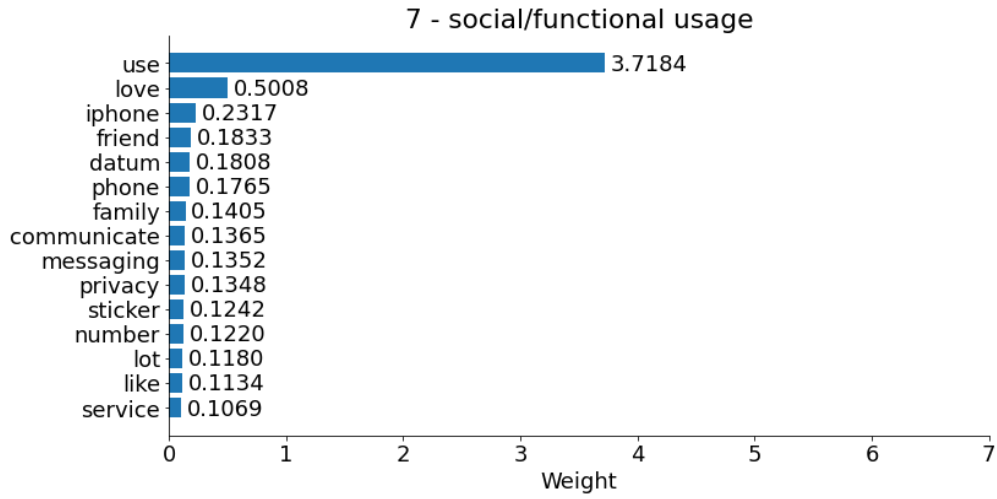
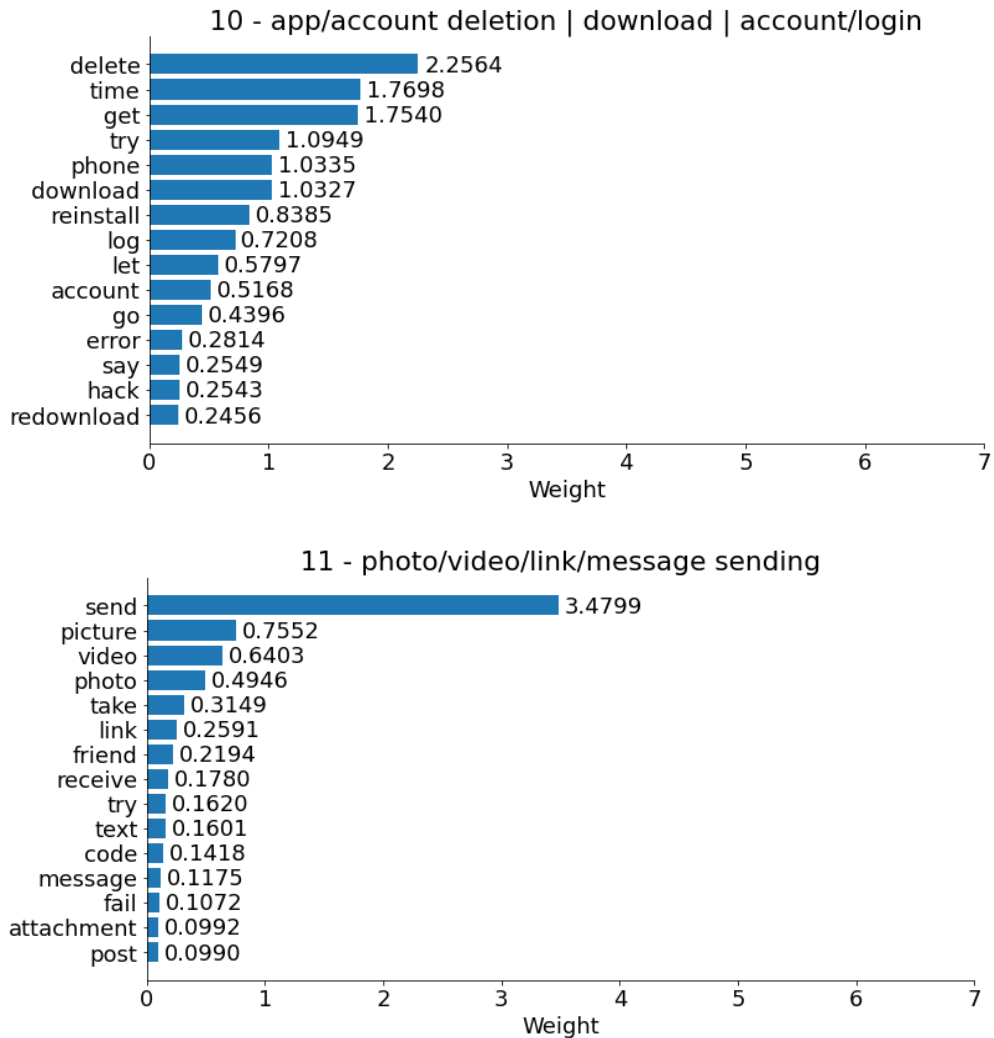


Figure 7 (Continued)



#### 4.1.1.2 Evaluation result of extracted topics

Table 10 shows the result of the manual evaluation in extracted topics with their accuracy scores sorted by topic in descending order. Based on the total number of 215 samples, an overall accuracy of 86.05% was achieved across all topics. The accuracy scores of eight topics (1 - app/link opening, 3 - update | ios version, 7 - social/functional usage, 5 - (video) call | connection quality, 11 - photo/video/link/message sending, 9 - general problems to fix, 0 - message/notification receiving | read receipt and 8 - status | search contact) are above the overall score. In detail, both topic 1 - app/link opening and topic 3 - update | ios version received a 100% accuracy, while topic 5 - (video) call | connection quality and topic 11 - photo/video/link/message sending received the same accuracy of 93.75%, topic 7 - social/functional usage in between, with 94.74%. Topic 9 - general problems to fix also obtained a score above 90%. In addition, the accuracy scores of four topics (0 - message/notification receiving | read receipt, 8 - status | search contact, 2 - (group) chat | add feature | privacy and 4 - app crash) fall between 80% and 90%,

while topic 10 - *app/account deletion | download | account/login* received the lowest accuracy (68.97%), about 2.5% lower than that of topic 6 - *app/feature working issue | apple watch*.

**Table 10**

*Accuracies of topic extraction*

Topic	# of samples	#. of incoherence	Accuracy
OVERALL	215	30	0.8605
1 - app/link opening	16		1.0000
3 - update   ios version	14		1.0000
7 - social/functional usage	19	1	0.9474
5 - (video) call   connection quality	16	1	0.9375
11 - photo/video/link/message sending	16	1	0.9375
9 - general problems to fix	14	1	0.9286
0 - message/notification receiving   read receipt	10	1	0.9000
8 - status   search contact	8	1	0.8750
2 - (group) chat   add feature   privacy	49	9	0.8163
4 - app crash	10	2	0.8000
6 - app/feature working issue   apple watch	14	4	0.7143
10 - app/account deletion   download   account/login	29	9	0.6897

#### 4.1.2 Evaluation result of sentiment analysis

Based on a total number of 214 samples, the confusion matrices (Table 11, Table 12, Table 13) present the statistical outcomes of the sentiment evaluation on weighted sentiment scores, user ratings and normalized VADER compound scores respectively. According to the manual labels, the majority class is *Negative* in the sample reviews. The numbers of correctly classified labels for weighted sentiment scores, user ratings and normalized VADER compound scores are 202, 192 and 182 respectively. As for the incorrectly classified labels, the number of false *Negative* (7) is slightly greater than that of false *Positive* (5) for weighted sentiment scores. On the contrary, the number of false *Negative* is obviously smaller than that of false *Positive* for user ratings and normalized VADER compound scores. For user ratings, the number of false *Positive* is 20 and of false *Negative* is only 2. For normalized VADER compound scores, the number of false *Positive* is 22 and of false *Negative* is 10.



**Table 11***Confusion matrix of sentiment evaluation on weighted sentiment scores*

	Weighted sentiment scores	
	Negative (< 4)	Positive (>= 4)
Negative (manual)	170	5
Positive (manual)	7	32

**Table 12***Confusion matrix of sentiment evaluation on user ratings*

	User ratings	
	Negative (< 4)	Positive (>= 4)
Negative (manual)	155	20
Positive (manual)	2	37

**Table 13***Confusion matrix of sentiment evaluation on normalized VADER compound scores*

	Normalized VADER compound scores	
	Negative (< 4)	Positive (>=4)
Negative (manual)	153	22
Positive (manual)	10	29

The calculated Precisions, Recalls and F1-scores for weighted sentiment scores, user ratings and normalized VADER compound scores are reported in [Table 14](#). Generally, the weighted sentiment scores outperformed the other two scoring methods, with an improvement of at least 7% in F1-score. Although the highest Recall (94.87%) was achieved when considering only user ratings as the sentiment scores, the Precision was about 30% lower than the Recall, resulting in a reduced F1-score. As for the normalized VADER compound scores, the Precision, Recall and F1-score were the lowest.

**Table 14***Performance of sentiment analysis*

	Precision	Recall	F1-score
Weighted sentiment scores	0.8649	0.8205	0.8421
User ratings	0.6491	0.9487	0.7708
Normalized VADER compound scores	0.5686	0.7436	0.6444

### 4.1.3 Results of competitive analysis

This sub-section presents the summary of review distributions and average sentiment scores and their corresponding visualizations as well as the result of sentiment evolution.

#### 4.1.3.1 Review distributions and average sentiment scores

The review distributions presented by percentages based on the review counts for each app ([Table 15](#)) are demonstrated in [Figure 8](#), [Figure 9](#), [Figure 10](#) and [Figure 11](#).

The average sentiment scores during the period between June 1, 2020 and May 31, 2021 were mainly separated into two score levels, above 3 and below 3 ([Table 15](#)). Overall, Signal received the highest overall average sentiment score of 3.6500, followed by Telegram with 3.2373. Contrarily, both Messenger and WhatsApp obtained overall average sentiment scores below 3, with 2.3108 for Messenger and 2.9090 for WhatsApp. As for the average sentiment scores by topic, none of the scores exceeded 3 for Messenger, and WhatsApp obtained scores below 3 for most of the topics. Signal and Telegram received obviously higher average sentiment scores for all topics. Ten out of twelve topics of Signal had average sentiment scores above 3, and average sentiments scores of seven topics for Telegram were above 3.

**Table 15**

Summary of review distributions and average sentiment scores

Topic	Messenger		WhatsApp	
	Count	Avg. score	Count	Avg. score
0 - message/notification receiving   read receipt	643	2.1940	182	2.9704
1 - app/link opening	1,231	2.2490	245	2.5101
2 - (group) chat   add feature   privacy	1,687	2.4907	1,955	3.1029
3 - update   ios version	830	2.2532	406	2.6615
4 - app crash	644	2.1761	264	2.7212
5 - (video) call   connection quality	461	2.5468	768	3.0304
6 - app/feature working issue   apple watch	785	2.2822	400	2.9034
7 - social/functional usage	690	2.3623	701	3.1048
8 - status   search contact	136	2.5203	523	2.8720
9 - general problems to fix	889	2.3479	359	2.6551
10 - app/account deletion   download   account/login	1,742	2.2118	736	2.5177
11 - photo/video/link/message sending	924	2.2795	398	2.9780
ALL TOPICS	10,662	2.3108	6,937	2.9090

Topic	Signal		Telegram	
	Count	Avg. score	Count	Avg. score
0 - message/notification receiving   read receipt	94	3.3207	48	3.3281
1 - app/link opening	29	2.9649	45	2.7441
2 - (group) chat   add feature   privacy	790	3.9519	499	3.5239
3 - update   ios version	62	3.3930	78	3.2011
4 - app crash	44	3.3079	34	2.9141
5 - (video) call   connection quality	202	3.4890	131	3.2824
6 - app/feature working issue   apple watch	155	3.7207	82	3.1065
7 - social/functional usage	300	3.9133	188	3.3475
8 - status   search contact	117	3.9722	56	3.8664
9 - general problems to fix	67	3.4091	82	2.8998
10 - app/account deletion   download   account/login	205	2.9414	194	2.7296
11 - photo/video/link/message sending	200	3.1979	94	2.8510
ALL TOPICS	2,265	3.6500	1,531	3.2373

#### 4.1.3.2 Visual review distribution

The review distribution of Messenger differed from those of the other apps (Figure 8, Figure 9, Figure 10, Figure 11).

For Messenger (Figure 8), reviews of topic 10 - *app/account deletion | download | account/login* and topic 2 - *(group) chat | add feature | privacy*, with nearly the same proportions, accounted for almost

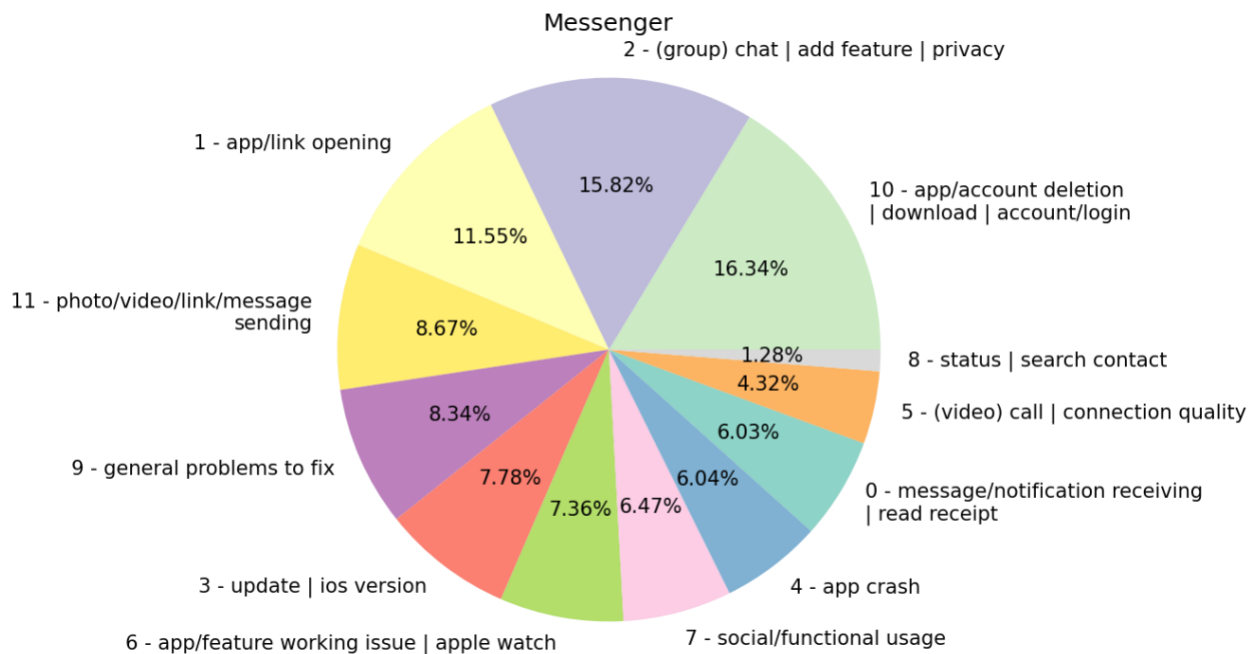
one third of its total reviews, followed by reviews of topic 1 - *app/link opening* (11.55%). Three topic groups, 11 - *photo/video/link/message sending* and 9 - *general problems to fix*, 3 - *update | ios version* and 6 - *app/feature working issue | apple watch* as well as 7 - *social/functional usage*, 4 - *app crash* and 0 - *message/notification receiving | read receipt*, respectively shared close percentages, around 8%, 7% and 6%. The percentage of topic 8 - *status | search contact* ranked last, with only 1.28%, following that of topic 5 - *(video) call | connection quality* (4.32%).

As for WhatsApp, Signal and Telegram (Figure 9, Figure 10, Figure 11), their distribution pattern shared a few similarities. Around one third of their respective reviews were related to topic 2 - *(group) chat | add feature | privacy*, while reviews of three topics, 5 - *(video) call | connection quality*, 7 - *social/functional usage* and 10 - *app/account deletion | download | account/login*, roughly took up another one third, although the exact percentages of these three topics varied from app to app. Also, the percentages of topic 0 - *message/notification receiving | read receipt*, topic 1 - *app/link opening* and topic 4 - *app crash* altogether occupied less than 10% for WhatsApp, Signal and Telegram.

In addition, reviews of topic 1 - *app/link opening* constituted 11.55%, which ranked third for Messenger (Figure 8). However, these percentages for WhatsApp, Signal and Telegram were about the least, with 3.53%, 1.28% and 2.94% respectively (Figure 9, Figure 10, Figure 11). Conversely, the percentage of topic 5 - *(video) call | connection quality* was 11.07% for WhatsApp (Figure 9), but that for Messenger was 4.32% (Figure 8), about only half of the corresponding percentage for Signal and Telegram (Figure 10, Figure 11).

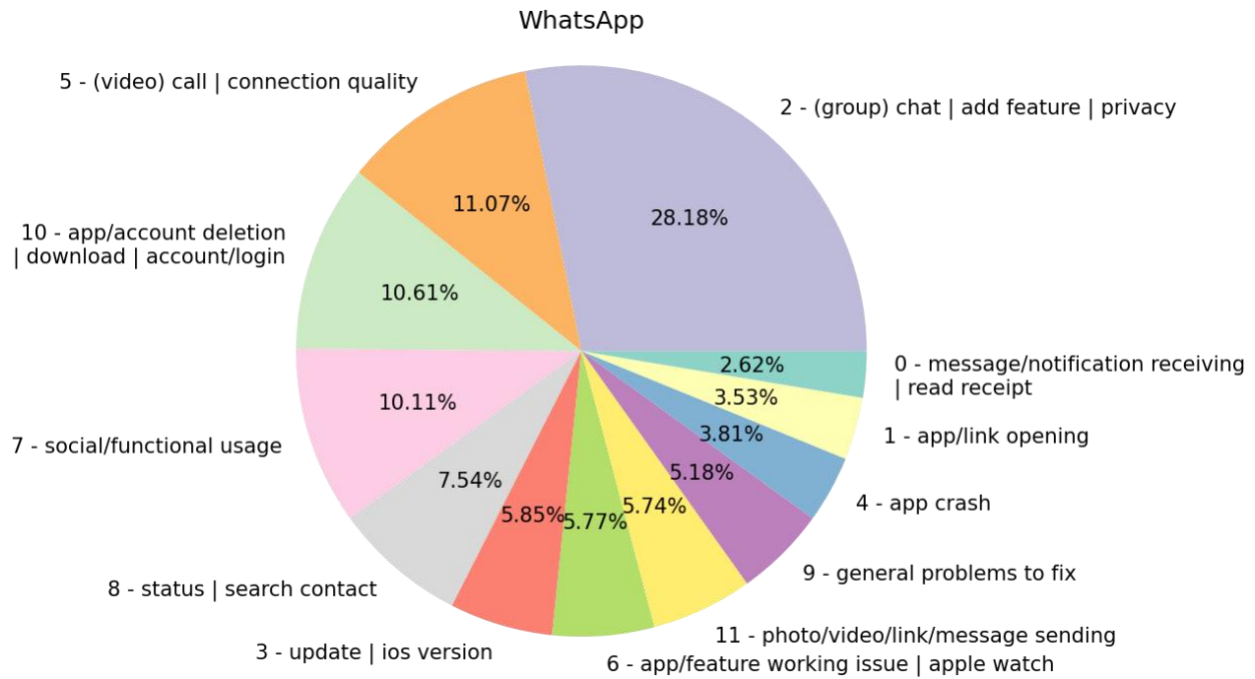
**Figure 8**

*Review distribution by topic of Messenger*



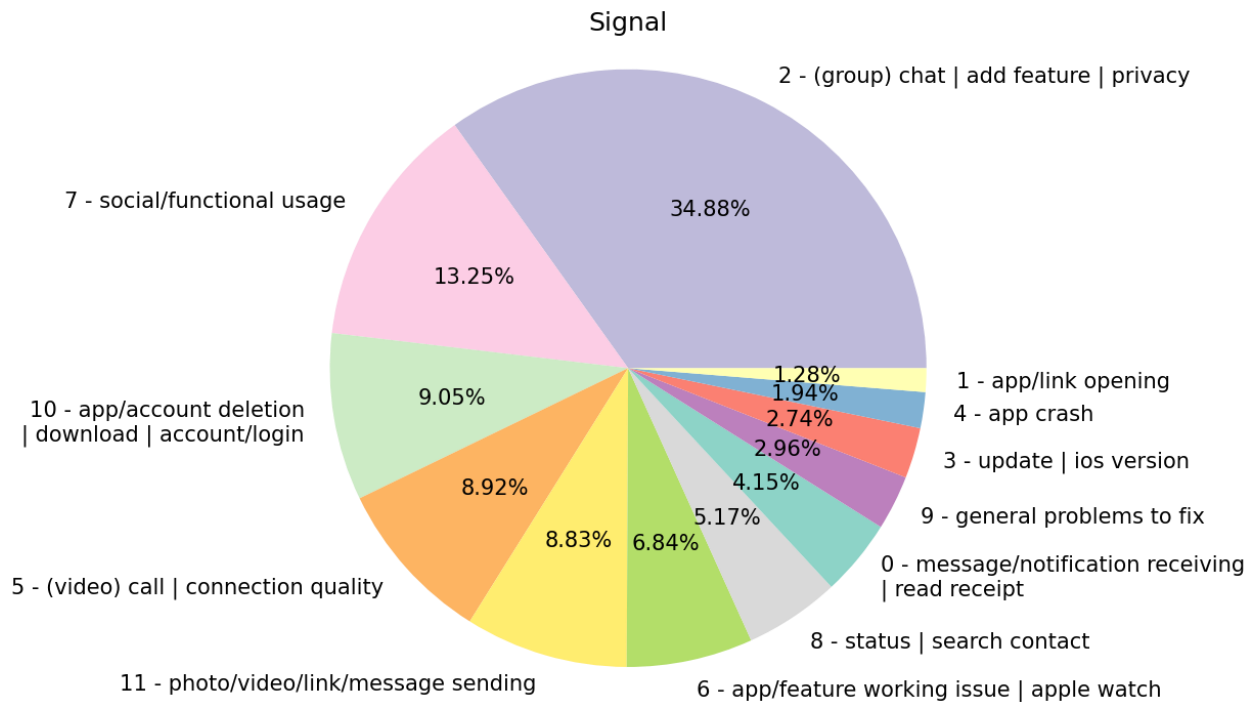
**Figure 9**

*Review distribution by topic of WhatsApp*



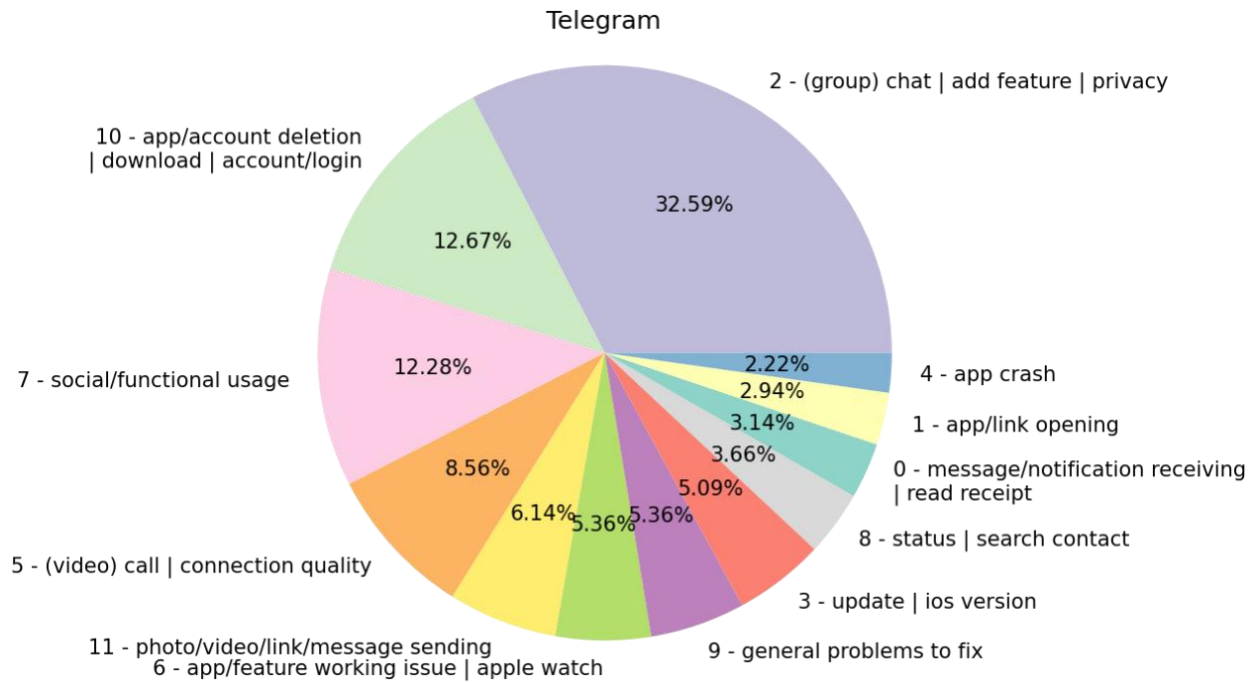
**Figure 10**

*Review distribution by topic of Signal*



**Figure 11**

*Review distribution by topic of Telegram*

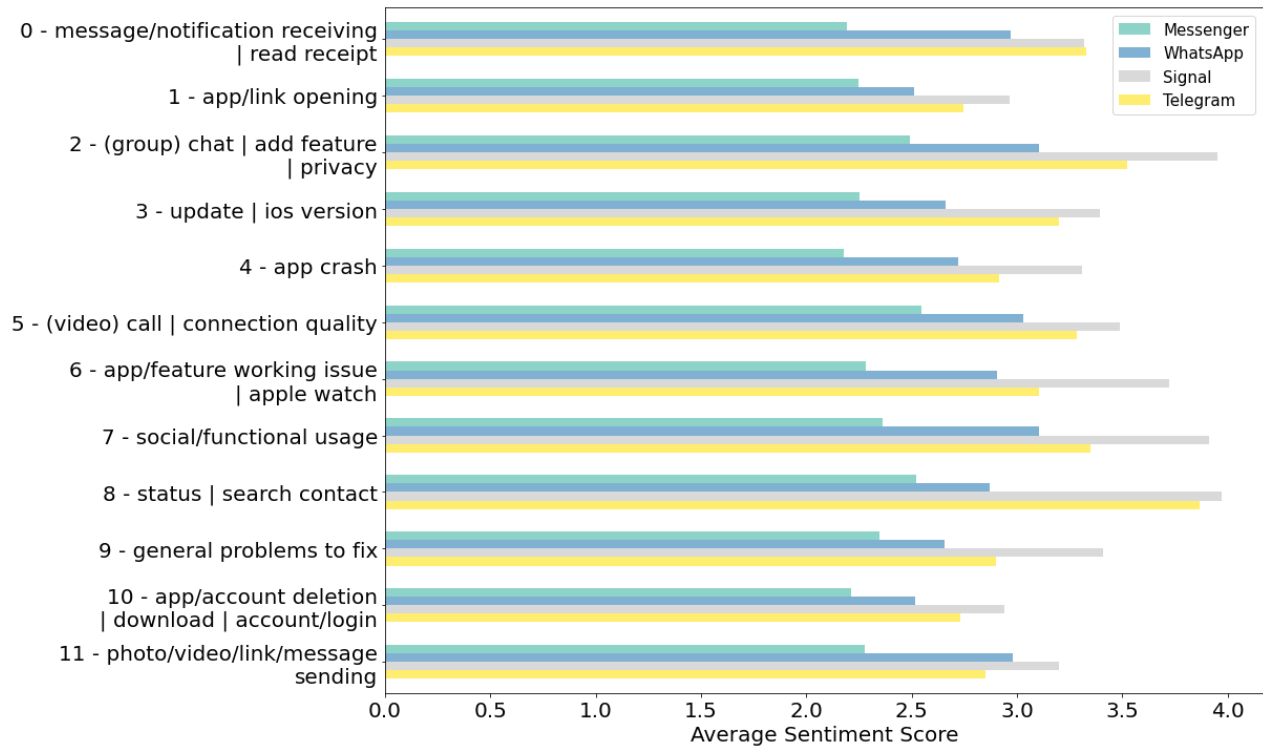


#### 4.1.3.3 Visual comparison of average sentiment scores

Figure 12 compares the average sentiment scores by topic for Messenger, WhatsApp, Signal and Telegram. The average sentiment scores of Signal were leading in nearly all the topics, except for topic 0 - message/notification receiving | read receipt, whose score was slightly lower than Telegram. Contrarily, Messenger obtained the lowest average sentiment scores in every topic. Although the average sentiment scores of Telegram generally outnumbered those of WhatsApp, WhatsApp received a higher score in topic 11 - photo/video/link/message sending. All apps obtained relatively lower average sentiment scores in topic 1 - app/link opening, topic 10 - app/account deletion | download | account/login and topic 11 - photo/video/link/message sending compared with other topics. The score gaps in topic 2 - (group) chat | add feature | privacy, topic 6 - app/feature working issue | apple watch, topic 7 - social/functional usage and topic 8 - status | search contact between Signal and Messenger were exceptionally large, with a difference about 1.5 points. Also, both Signal and Telegram received an obvious higher average sentiment score than Messenger and WhatsApp in topic 8 - status | search contact.

**Figure 12**

*Comparison of average sentiment scores by topic*



#### 4.1.3.4 Sentiment evolution

Figure 13 demonstrates the monthly changes in average sentiment scores of each topic between June 2020 to May 2021 for Messenger, WhatsApp, Signal and Telegram based on the statistical aggregation of average sentiment scores by month, topic and app (Table 16). The sentiment evolution of Messenger and WhatsApp in all topics was complete, while Signal and Telegram had some missing records in specific topics. Signal had missing records in topic 0 - message/notification receiving | read receipt, topic 1 - app/link opening, topic 3 - update | ios version, topic 4 - app crash, topic 6 - app/feature working issue | apple watch and topic 8 - status | search contact. The average sentiment scores of topic 8 - status | search contact, in particular, were missing from July 2020 to November 2020. Telegram had missing scores in topic 4 - app crash and topic 8 - status | search contact.

The average sentiment scores of Messenger remained stable at a low level for all topics during the time. A similar sentiment evolution pattern with a slightly higher level of scores was observed for WhatsApp, although WhatsApp experienced a temporary fluctuation in topic 0 - message/notification receiving | read receipt around August 2020 and in topic 7 - social/functional usage around Jan 2021. By contrast, the scores of Signal and Telegram were not so stable as Messenger and WhatsApp in nearly all topics, with a dramatic fluctuation especially in topic 1 - app/link opening, topic 3 - update | ios version and topic 4 - app crash. Although the average sentiment scores of Signal generally stayed at the highest level for all topics in comparison with those of the other apps, lowest scores were recorded for topic 1 -

*app/link opening* in October 2020 and for topic 4 - *app crash* in July, September and October 2020. As for Telegram, the lowest average sentiment score for topic 3 - *update | ios version* in April 2021 and for topic 11 - *photo/video/link/message sending* during the period between February to April 2021 were recorded. Also, the average sentiment scores of topic 0 - *message/notification receiving | read receipt*, topic 5 - *(video) call | connection quality* and topic 8 - *status | search contact* suffered a mild downward tendency in fluctuation since November 2020.

**Figure 13**

*Sentiment evolution by topic*

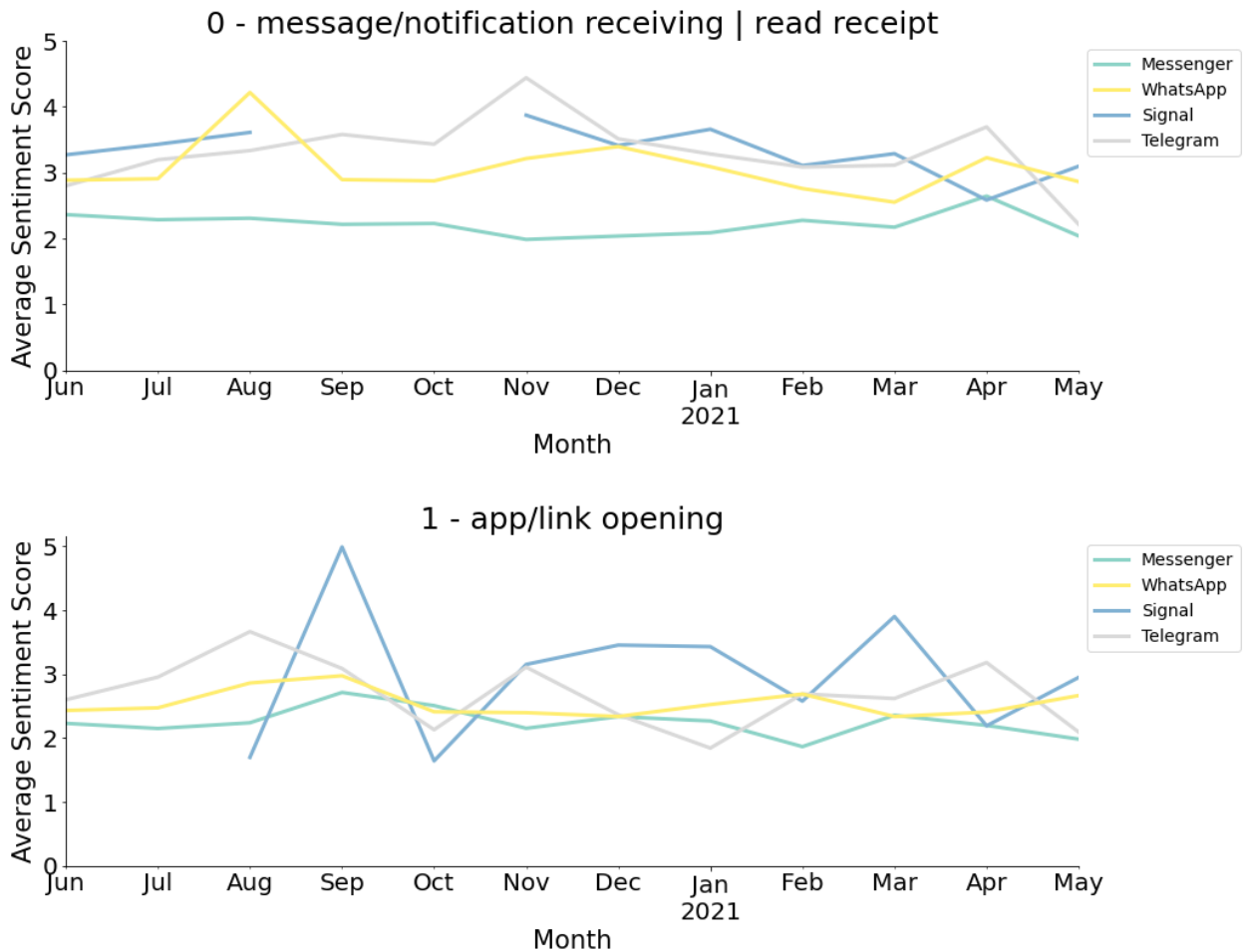




Figure 13 (Continued)

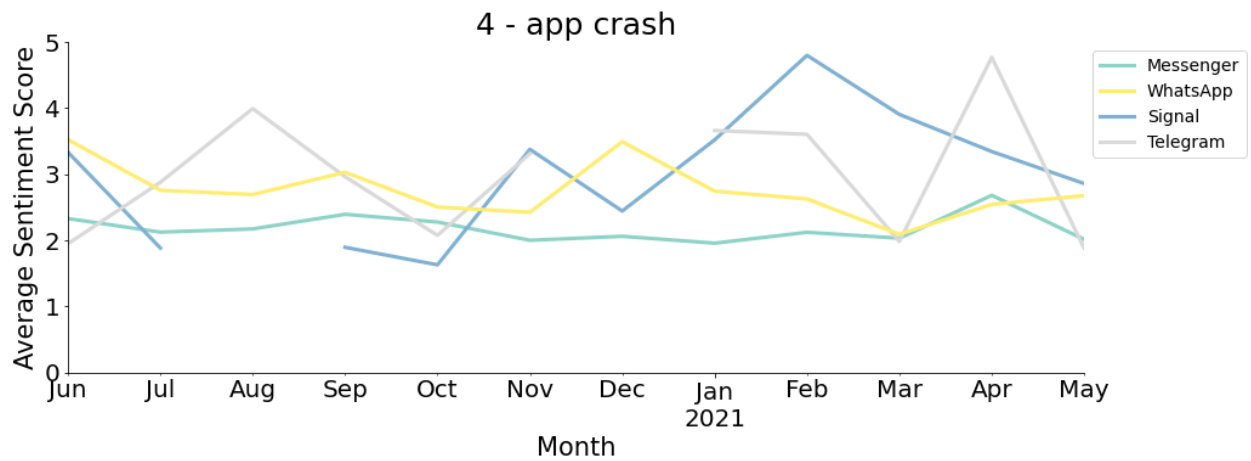
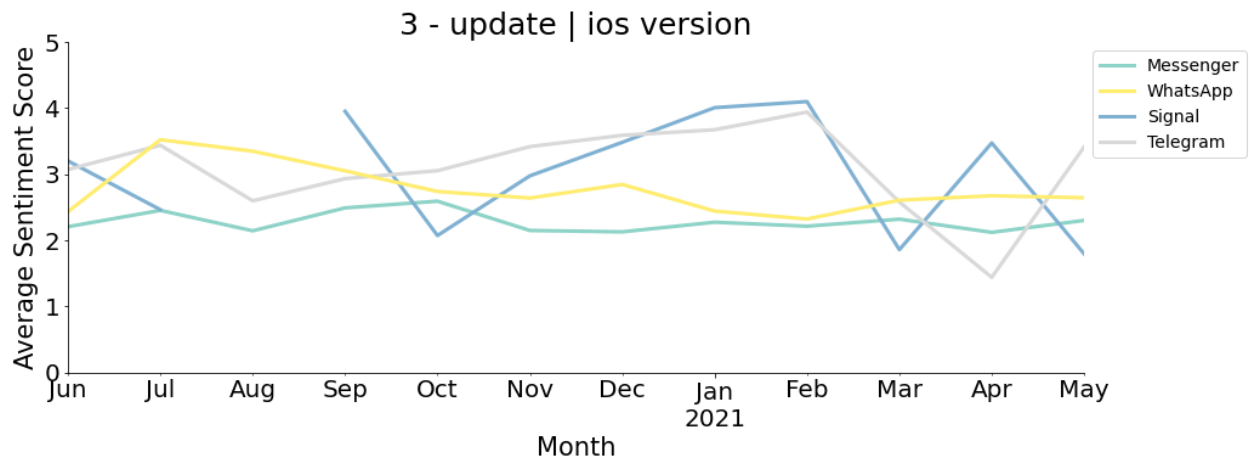
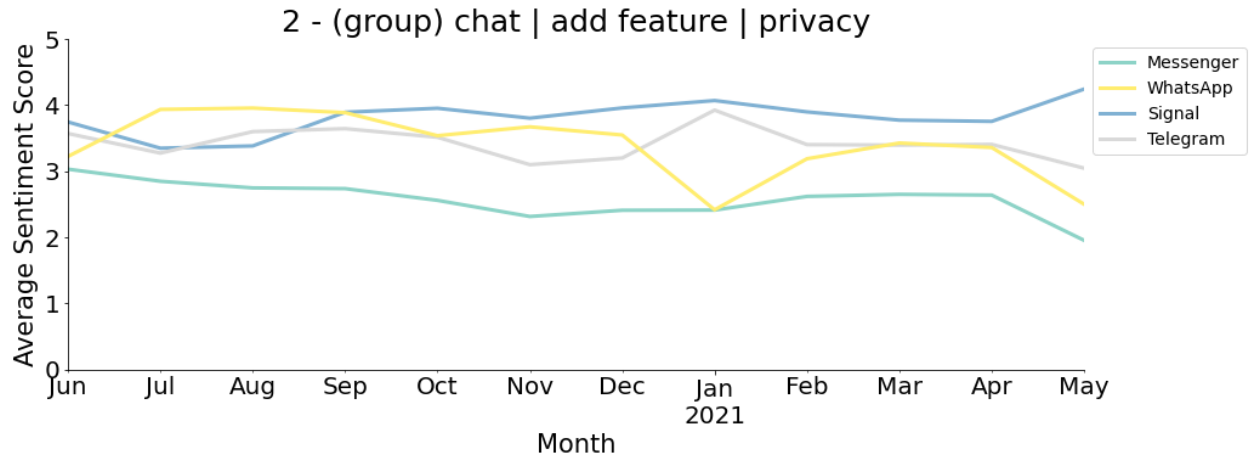


Figure 13 (Continued)

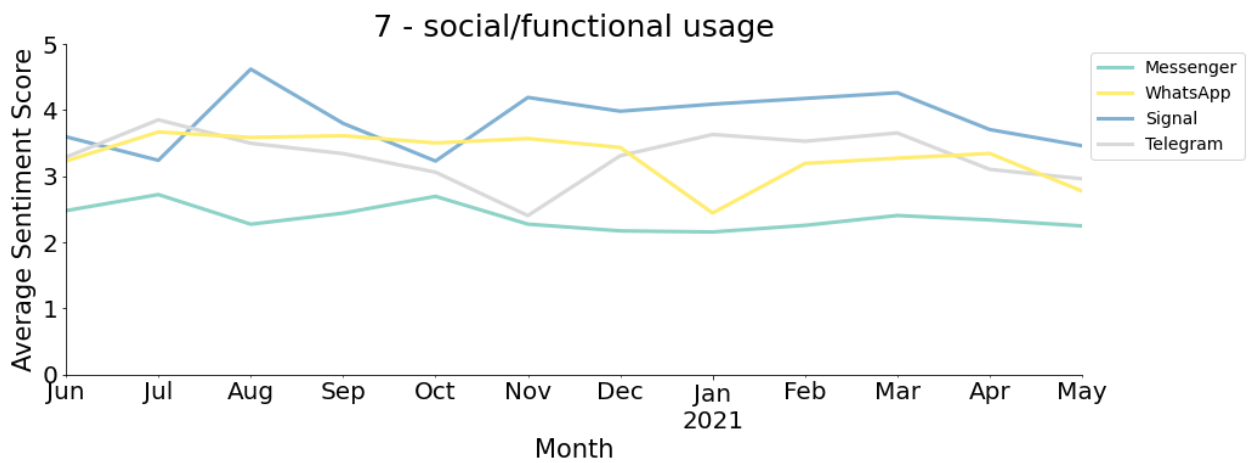
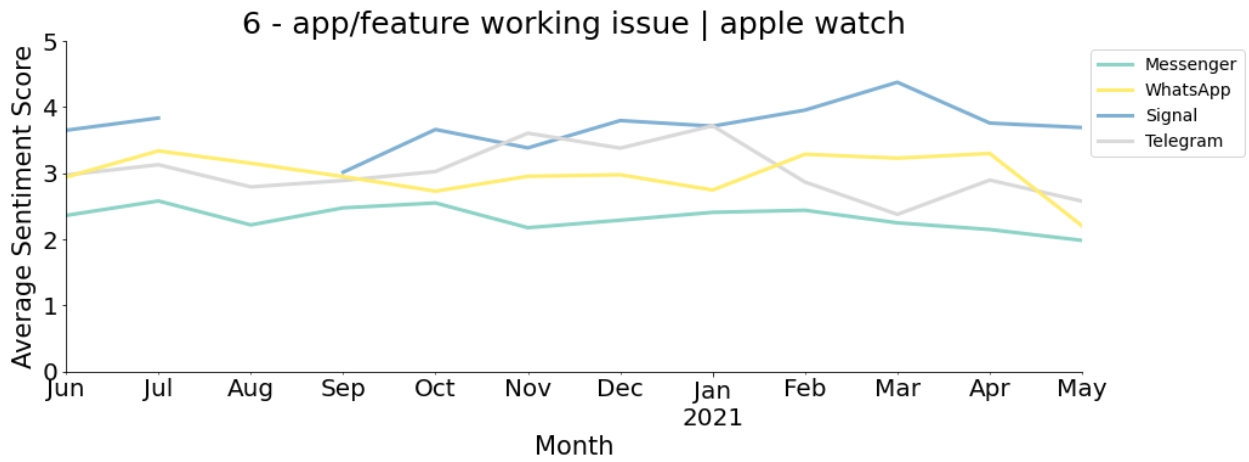
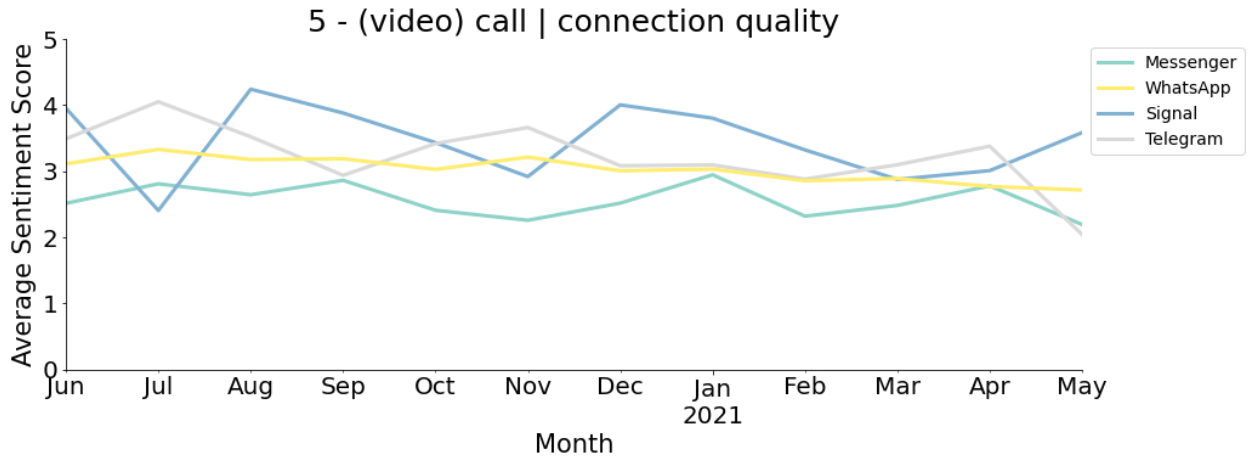


Figure 13 (Continued)

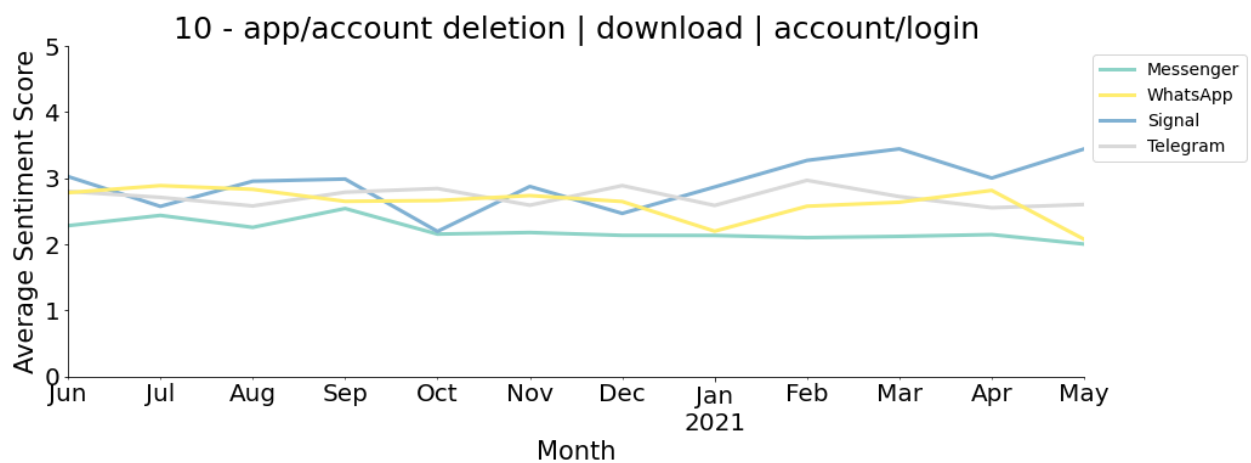
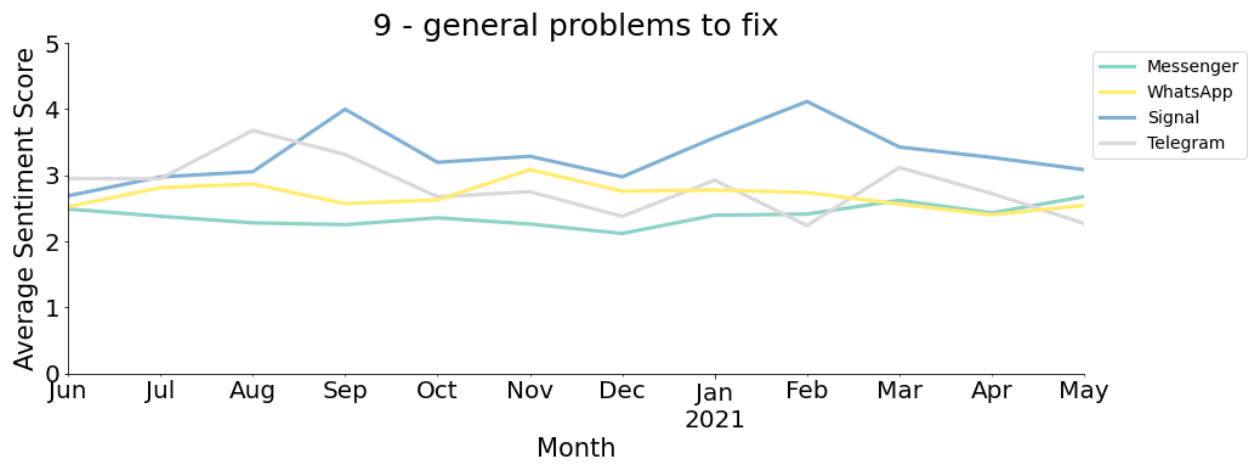
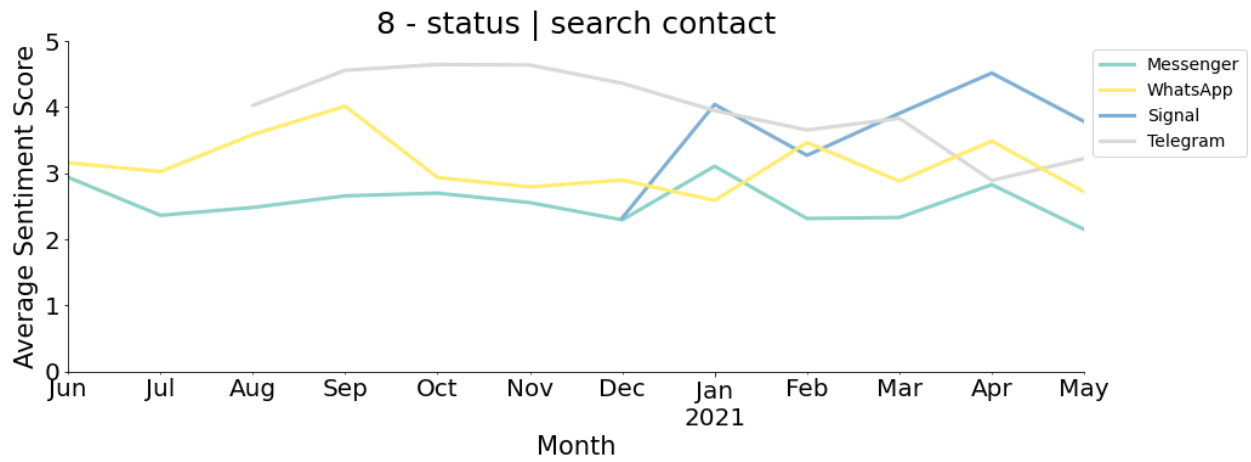
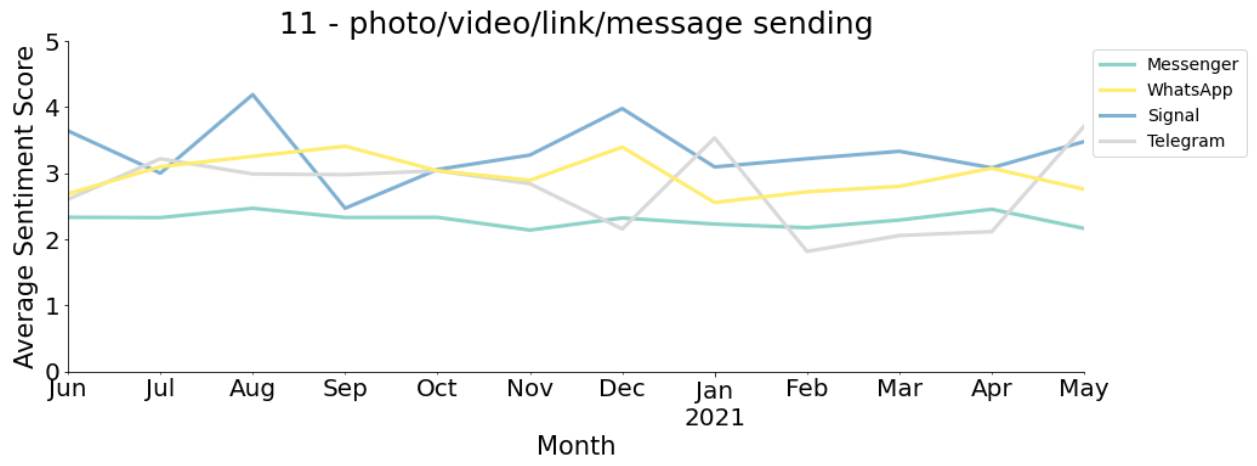


Figure 13 (Continued)





## 4.2 DISCUSSION

By text-mining 27,479 user reviews of four messaging apps released on Apple Store, the present work combined topic modeling and sentiment analysis to perform competitive analysis. The results show that the topic model extracted comprehensible topics with promising accuracies, and that the sentiment analysis performed well. Additionally, the competitive analysis based on the extracted topics and the weighted sentiment scores revealed meaningful competitive intelligence in terms of several facets.

### 4.2.1 Topic modeling

In the present work, the accuracies ranging from 69.97% to 100% in topic extraction suggest the NMF topic model works well with user reviews of messaging apps. The relatively lower accuracies might partially result from a fair number of reviews that express protests against a recent attack in Palestine. The language expression of political discussion is likely to confuse the topic decomposition of the model. Moreover, we followed the approach adopted by [Vu et al. \(2015\)](#) to use nouns and verbs as app features and obtained an improvement of 2.94% in overall accuracy compared with their average accuracy of 83.11% using keyword-based approach to analyze user reviews from Google Play. This improvement might be due to a different approach to mine the app aspects or a more domain-specific text preprocessing for messaging apps. Another possible explanation can be that the expression pattern of user reviews from Google Play is different from that from Apple Store. Furthermore, [Guzman and Maalej \(2014\)](#) used the LDA approach to extract topics from collocations of nouns, verbs and adjectives in user reviews for multiple non-competitive Android and iOS apps including WhatsApp. Their results show that the precisions and recalls varied on an app basis, and the highest precision and recall were achieved for WhatsApp, with F1-scores of 0.781 and 0.813 respectively for inclusion and exclusion of sentiment words in feature words. These F1-scores for WhatsApp in their result, combining the overall accuracy in topic extraction in our result, might indicate that topic models are particularly effective to extract app features from user reviews of messaging apps but not any type of mobile apps. A possible explanation can be that messaging apps usually have homogeneous functionalities, and thus users of such apps might tend to express a specific app aspect in similar ways. Also, their exclusion of sentiment words provides a feasible way to improve our performance of topic extraction by pruning sentiment words in app feature terms. In their study, the sentiment words are usually adjectives (e.g., great, bad) and verbs (e.g., hate, love), while in our research, feature terms that bear sentiments are mostly verbs.

There are some important findings in the extracted topics. Firstly, domain-specific knowledge was shown in some of the topic names. The term “app” is not in any of the topic words in topic 1- *app/link opening*, topic 4 - *app crash*, topic 6 - *app/feature working issue | apple watch* or topic 10 - *app/account deletion | download | account/login*, but their topic names still relate to app due to the logical context of their topic words. In fact, it was impossible to have the term “app” in any list of topic-words since this term was expanded to “application” and then removed as a stop word due to its extremely high frequency in the corpus. Secondly, topic-words such as “uninstall” and “reinstall” appeared in several topics such as topic 1 - *app/link opening*, topic 4 - *app crash* and topic 6 - *app/feature working issue | apple watch*. One possible explanation can be the user behavior of uninstalling and reinstalling the app for attempting to solve an issue, e.g., app crash, and the users mentioned such behaviors along with

other app feature issues in the reviews. Thirdly, a few similarities in topic keywords were found between some extracted topics in the present work and some topics in Su et al.'s (2019) study. They applied the LDA method to find latent topics from user reviews of four mobile apps on Google Play, and Messenger was one of those apps. Their study lists some extracted topics with top five keywords for Messenger. As illustrated in Table 17, some of their top five keywords can also be found in the top fifteen keywords of three extracted topics in the present work. It is possible that more topic-words could be matched if they had shown more keywords of each topic, e.g., top fifteen keywords rather than only top five keywords. The app aspects discussed in each pair of the three pairs of topics are very close based on the logical connection of the keywords in each topic. One explanation for this could be that nearly half of the total user reviews for topic modeling are of Messenger in the present work (Table 15). It is also likely that iOS users and Android users tend to discuss similar app aspects or report similar app issues of messaging apps in user reviews. Lastly, some of the extracted topics comprise multiple incoherent aspects. For instance, in topic 2, the three aspects of *(group) chat*, *add feature* and *privacy* could be three individual topics. This topic incoherence might result from a significant number of reviews mentioning the three aspects together. Another explanation could be that the number of topics  $k$  was defined to be too small, and thus the topic model generated excessively broad topics. Similar problem of the number of  $k$  might also happen to topic 7 - *social/functional usage*. This topic required a comprehensive consideration of other words with smaller weights and a combination with its abstract primary word "use" for condensing the main issues discussed in that topic. The words such as "love", "friend", "family" and "communicate" suggest social aspects, while the words like "iphone", "datum", "phone" and "messaging" indicate functional features. However, social and functional usage is an overly broad topic, which can refer to many app aspects. Overall, the extracted topics suggest that defining the number of  $k$  is a major challenge, but the NMF topic model is still useful for extracting topics, most of which reflect clear and separable app aspects of messaging apps.

**Table 17**

*Comparison of keywords in extracted topics*

The present work - NMF Messenger, WhatsApp, Signal, Telegram		Su et al. (2019) - LDA Messenger	
Topic id	Top 15 keywords	Topic id	Top 5 keywords
0	<b>message, notification</b> , show, get, see, receive, read, go, <b>request</b> , say, check, voice, type, reply,	3	<b>message, notification</b> , delete, give, <b>request</b>
2	<b>chat</b> , see, people, <b>feature, add</b> , want, love, option, <b>group</b> , need, privacy, make, friend, <b>change</b> , go	8	<b>chat, change, feature, add, group</b>
5	<b>call</b> , video, <b>voice</b> , phone, make, quality, <b>connect</b> , hear, ring, receive, audio, drop, issue, play, sound	0	<b>call, screen, turn, voice, connect</b>

### 4.2.2 Sentiment analysis

Through a comparison of different sentiment scoring methods, we discovered the weighted sentiment scores mitigated the bias of user ratings and benefited from the normalized VADER compound scores, resulting in a relatively objective sentiment scoring for further competitive analysis.

Intriguingly, [Su et al. \(2019\)](#) conducted a similar sentiment analysis using average weighted scores for two groups of similar apps released on Google Play and obtained on average a precision of 92.29%, a recall of 71.41% and a F1-score of 80.24%. Their precision-recall tradeoff is contrary to those of user ratings and normalized VADER compound scores in our result, which tended to receive lower precisions and higher recalls ([Table 14](#)). This might be due to the different score benchmarks for separating positive and negative reviews, 2.5 in their study but 4 in the present work, producing a majority class of negative in our sentiment classification. Comparatively, a score benchmark of 4 seems more appropriate for separating the sentiment polarity for user reviews of mobile apps. Our result shows that the weighted sentiment scores based on this score benchmark obtained a better F1-score with a balanced precision-recall tradeoff.

Furthermore, in the sentiment analysis of the present work, we averaged the weighted sentiment scores of user reviews by topic as the topic sentiment score, while in [Su et al.'s \(2019\)](#) study, the sentiment score for each topic found by the LDA model takes the number of users' hitting the like button into consideration. However, there are not only users who click the like button but also users who click the dislike button. It is possible that the count of user dislikes outnumbers the count of user likes for specific reviews. Merely considering the number of user likes might bias the sentiment score of a topic. In spite of this possible bias, their study still encourages a more elaborate sentiment analysis, which could take both the number of user likes and the number of user dislikes into account when it comes to the calculation of the sentiment score of a topic.

### 4.2.3 Competitive analysis

The competitive analysis revealed meaningful insights into user concerns, competitive strengths and weaknesses as well as changes of user sentiments over time. According to [Porter \(1980\)](#), the components of competitor analysis comprise multi-dimensional objectives of the competitor's managerial personnel, assumptions for the competitor itself and for the industry, competitive strategy and the competitor's resources and capabilities including strengths and weaknesses. Thereinto, the objectives and assumptions reveal what drives the competitor, while the strategy, resources and capabilities reflect what the competitor is doing or is capable of doing. Based on these components of competitor analysis, the user concerns, competitive strengths and weaknesses as well as changes of user sentiments over time shown in the present work meet what the competitor is doing or is capable of doing. Specifically, user concerns, on the one hand, reflect in a way the capabilities of a competing messaging app, and on the other hand, user concerns also provide legitimate ground for conjecturing the competitive strategy of the rival. Also, competitive strengths and weaknesses as well as changes of user sentiments over time uncover the resources and capabilities of a competitor.



The review distributions by topic for the four messaging apps reflects the major user concerns about specific app aspects. In general, app users were very concerned about chat features and privacy issues. Their attention to chat features conforms to people's perception mainly because chat features are the core functions of all messaging apps. Privacy issues should arouse the attention of practitioners in the industry of instant messaging services. Data security and privacy may become a key competitive advantage for differentiating one messaging app from the others. Another major user concentration is the download and account issues ranging from download, login problem and account deletion. These issues are usually not related to the key features of apps but might be the generic problems of many mobile apps, which bother the app users very much. Solving these problems by better functional designs and technical supports would improve the user satisfaction. For example, after downloading, users could choose multiple ways to log in the app without any login failure. Moreover, users of WhatsApp, Signal and Telegram had similar focuses on the app aspects particularly regarding voice and video call, connection quality, social and functional usage, account and download issues. Aspects such as voice call and video call, just like the chat features, are also the crucial features of apps for daily communication. Their users might do frequent voice or video call apart from text messages, and thus they attached great importance to the quality of calls. An excellent and stable connection quality would be a big attraction for users who make lots of voice calls or video calls. Also, these three apps might be very important for their users to keep in touch with their family and friends. During the social communication including text messages, voice calls and video calls, the functional usage, e.g., mobile data usage, might draw the special attention of app users. It is likely that a specific feature of mobile data saving mode in the app settings can become a unique competitive advantage in the era of mobile Internet. As for Messenger, in addition to the chat features, privacy, download and account issues, the problems regarding app and link opening were also mentioned by many users. Nevertheless, this aspect was not highly discussed by users of WhatsApp, Signal and Telegram. This contrast might suggest that users of Messenger suffered from more failure or inconvenience in opening the app and the links in messages. Regarding the app opening, the problem might be no reaction after clicking on the app icon. The problem of app and link opening deserves the attention of their research and development personnel, who might further investigate the specific issues by retrieving user reviews of that topic. Another interesting finding is that users of Messenger did not discuss the aspect of voice call, video call and connection quality as much as the users of WhatsApp, Signal and Telegram. One possible explanation can be that the users of Messenger did not encounter outstanding issues or frequently express their emotions about this app aspect. It could also be that users of Messenger mainly texted messages rather than made voice calls or video calls, and thus they did not pay much attention to this app aspect. In the latter situation, the reasons can be further examined. The voice and video call of Messenger might be inconvenient to use, and the user might need additional clicks to use this app feature. Lastly, aspects of the messaging apps such as general problems to fix, app and feature working issues as well as app crash mainly pertain to technical issues. Other app aspects such as messaging sending and receiving, read receipt, status, contact searching and update were also discussed by users of these messaging apps. These app aspects might relate to not only technical issues but also design of the apps. The specific solutions to these problems should be determined based on the specific issues mentioned in the user reviews. The app aspect-based topics could provide the practitioners of messaging apps with clearer directions for user concerns and troubleshooting.

The comparison of average sentiment scores by app aspect-based topic reveals the competitive strengths and weaknesses of the four messaging apps. The higher average sentiment scores in nearly all app aspects obtained by Signal and Telegram might suggest that these two apps were leading the new trend of messaging apps. Both existing players and new entrants in the market might be able to get useful inspirations from their app designs and company philosophy. This speculation based on the distinct user sentiments of different messaging apps is not unfounded. The results of [Kim and Jeong's \(2018\)](#) study indicate that the sentiment portrayed by the opinions of users clearly distinguished the market leader from the market follower, even though they applied the sentiment analysis to online UGC for a different industry, two competing ramen brands in Korea rather than messaging apps. Their study provides compelling support for the validity of user sentiments in revealing competitive intelligence. In our research, market leaders and market followers were not necessarily differentiated from user sentiments because a market leader usually suggests a player with the highest market share in the industry. For messaging apps, varying release dates of the apps resulted in disparate user bases. Messenger received the most user reviews among the four messaging apps during the period between June 1, 2020 and May 31, 2021, and it probably owns the largest user bases compared with the other three messaging apps. However, Messenger received the lowest average sentiment scores in every app aspect-based topic. Hence, in our research, user sentiments tend to reveal the performance of specific app aspects and the user preference towards specific app aspects. From this point of view, our speculation about the new trend of messaging apps is reasonable to a certain extent. Based on this speculation, it is worth noting that WhatsApp, although generally behind Signal and Telegram, was still competent in the app aspect of messaging sending, including the sending of photos, videos, links and text messages. WhatsApp might have a user-friendly design of this app feature, which possibly facilitated its users to send messages in various forms. Also, with generally positive user sentiments in almost every app aspect, Signal and Telegram far outperformed Messenger and WhatsApp in the status feature and the contact searching. The design of the online and last seen status as well as the contact searching of these two apps might be very different but probably cater for many users in varying degrees. The configuration of the online and last seen status might be very flexible, or this feature might not even exist in consideration of privacy. The design of the contacts might be very user-friendly with easy searching feature without contact loss when users log in to the app account on another phone. Furthermore, all of the four messaging apps obtained relatively lower average sentiment scores in app aspects regarding message sending, app and link opening, download and account issues. The former two app aspects are of the most important features of messaging apps, and thus app users might emphasize their negative emotions when they encounter problems in sending messages of any forms and opening the apps or links. As for the download and account issues which were previously discussed in the review distribution by topic, these issues might not only draw the extra attention of app users, but also negatively impact user satisfaction. Overall, user sentiments of Signal were the most positive in almost every app aspect, while those of Messenger were the most negative in all app aspects. Particularly, the disparity of user sentiments focused on the app aspects regarding chat features, privacy, general working issues of app and features, apple watch issues, social and functional usage, status as well as contact searching. It is likely that the app design related to these aspects of Signal embodies significant advancement based on the corresponding app aspects of Messenger. Practitioners of messaging apps

might be able to get some valuable insights by further investigating the differences of the corresponding app aspects between these two apps.

The sentiment evolution indicates the changes of user sentiments over time of the four messaging apps in each app aspect-based topic. From the sentiment evolution, the user sentiments of Messenger and WhatsApp retained generally stable during the year. On the contrary, dramatic fluctuations were seen in the sentiment evolution of most app aspects for Signal and Telegram, particularly the app aspects related to app and link opening, version update as well as app crash. The sentiment evolution of Signal generally maintained at the highest level in all app aspects. However, Signal received lowest average sentiment scores in the app aspects of app and link opening as well as app crash in specific months. The similar situation also happened to Telegram, with lowest average sentiment scores in the app aspects of version update and message sending during some months. A possible explanation can be that they are newly released apps developed by smaller companies compared with Messenger and WhatsApp, which were developed by technology giants, and thus these two apps might suffer from more instabilities in technical support and maintenance. Additionally, the app aspects regarding app and link opening, version update as well as app crash with dramatic fluctuations in sentiment evolution might suggest that users of Signal and Telegram were more concerned about those app aspects. As a result, any issues regarding those aspects were more likely to trigger the emotional reactions of users. Moreover, user sentiments of Telegram in the aspects concerning message and notification receiving, read receipt, voice call and video call, connection quality, status as well as contact searching experienced moderate declines since November 2020. Telegram might make some modifications of these app features, which its app users probably did not prefer. Another interpretation could be that the users of Telegram had been suffering from constant technical issues such as bugs and crashes in these app aspects. Finally, both Signal and Telegram had some missing average sentiment scores of specific app aspects such as app crash, status and contact searching during the sentiment evolution. This might not be a bad circumstance probably because their users did not bring up any noteworthy problems regarding those app aspects during some months.

In general, the major user concerns of messaging apps might suggest the frequent issues highlighted by their users. To tackle these issues, these players of messaging apps may come up with relevant countermeasures. It would be advantageous for practitioners in the industry to have an anticipation of the potential competitive strategies of their rivals. Also, competitive strengths and weaknesses as well as changes of user sentiments over time of the competing apps enable practitioners in the same market to better understand the capability gaps and resource allocation of their competitors from the perspective of app users. To sum up, the competitive insights into these aspects may help existing players and new entrants in the market of messaging apps to enhance the comprehension of their position in the competitive landscape and grasp the development direction of the industry based on user needs.

## 5 CONCLUSION

To succeed in progressively fierce competition, practitioners of messaging apps need to possess a thorough knowledge of both their user needs and those of competitors. UGC has become a valuable source for understanding the real demands of users. Unlike industries such as hospitality and retail business, mobile apps have been rarely studied for the purpose of gaining competitive intelligence from user feedback.

We are thus motivated to perform a competitive analysis by combining topic modeling and sentiment analysis on user reviews of messaging apps. This work aims to examine the usefulness of topic modeling and sentiment analysis for revealing meaningful competitive intelligence from user reviews. With this objective, we employed the NMF topic model to find the latent topics regarding app aspects from user reviews and leveraged the VADER sentiment analysis tool as well as user ratings to adjust the sentiment score for each review after a domain-specific text preprocessing on user reviews of four messaging apps released on Apple Store. Thereafter, we conducted a competitive analysis using the extracted topics and the adjusted sentiment scores. The results show that the topic model properly found app aspect-based topics, and that the adjusted sentiment scores better represented the real user sentiments. Based on the outcome of the competitive analysis, it can be concluded that a combination of the NMF topic model and the adapted sentiment analysis is effective and useful for uncovering significant competitive insights into user concerns, competitive advantages and disadvantages as well as the development of user sentiments over time.

Compared with the extant studies regarding comparative analysis for mobile apps, the present work made full use of all informative user reviews rather than only reviews with comparative expressions and adopted a topic model to automatically extract app aspect-based topics across multiple competitive apps without any manual filter or additional topic matching. More importantly, we aggregated the extracted topics and the weighted sentiment scores to perform a competitive analysis, which was not further explored in most studies. We argued that this competitive analysis is highly critical to demonstrate the usefulness of the topic modeling and sentiment analysis that were implemented anteriorly. The insights revealed by the competitive analysis can help existing players and new entrants in the market of messaging apps at three levels: (1) understanding the principal concerns of users, (2) knowing the competitive gap based on strengths and weaknesses, and (3) detecting possible circumstances of competitors from the variation of their user sentiments. The knowledge of these three levels is expected to form the basis of corresponding strategies not only for prioritizing resource allocation but also for leading the development trend of the industry.

## 6 LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

This work has several limitations. Firstly, evaluations were performed manually by the author who is not working in the industry of instant messaging services. To alleviate the potential bias, we plan to invite practitioners in the industry to collectively evaluate the performance of topic modeling and sentiment analysis and interpret the results of competitive analysis. Secondly, the number of twelve topics to be extracted was defined based on an estimation of the possible number of topics, but different numbers can lead to distinct results in topic modeling. Extracted topics will be too general if the number of topics  $k$  is too small. In contrast, number of topics  $k$  being too large will generate many analogous or even overlapped topics. In the future, we will consider [Greene et al.'s \(2014\)](#) stability-based method for selecting the appropriate number of topics for topic modeling. Thirdly, only user reviews on the Apple App Store were analyzed. One important future direction is to adapt the present approach to user reviews on the Google Play. Lastly, we only text-mined reviews of users from the United States. However, users of messaging apps are spread all over the world, and thus user feedback might vary from region to region. To reveal more integrated competitive intelligence in instant messaging market, we plan to analyze more reviews of users from populous countries in Asia, Europe, South America and Africa. These reviews involve different languages and multilanguage text analytics will be the key challenge for future researchers.

## 7 BIBLIOGRAPHY

- Agarwal, D., & Chen, B.-C. (2010). fLDA: matrix factorization through latent dirichlet allocation. *WSDM '10: Proceedings of the third ACM international conference on Web search and data mining*, 91–100. <https://doi.org/10.1145/1718487.1718499>
- Akkaya, C., Wiebe, J., & Mihalcea, R. (2009). Subjectivity word sense disambiguation. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 190–199. Retrieved April 08, 2021, from <https://www.aclweb.org/anthology/D09-1020.pdf>
- Albalawi, R., Yeap, T. H., & Benyoucef, M. (2020). Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Front. Artif. Intell.*, 3, 42. <https://doi.org/10.3389/frai.2020.00042>
- Amadio, W. J., & Procaccino, J. D. (2016). Competitive analysis of online reviews using exploratory text mining. *Tourism and Hospitality Management*, 22(2), 193–210. <https://doi.org/10.20867/thm.22.2.3>
- Appfollow. (n.d.). *AppFollow: App Review Management & ASO Platform*. <https://appfollow.io/>
- Archak, N., Ghose, A., & Ipeiritos, P.G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science*, 57(8), 1485–1509. <https://doi.org/10.1287/mnsc.1110.1370>
- Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, 2200–2204. Retrieved April 08, 2021, from [http://www.lrec-conf.org/proceedings/lrec2010/pdf/769\\_Paper.pdf](http://www.lrec-conf.org/proceedings/lrec2010/pdf/769_Paper.pdf)
- Benamara, F., Cesarano, C., Picariello, A., Reforgiato, D., & Subrahmanian, V. S. (2007). Sentiment Analysis: Adjectives and Adverbs are Better than Adjectives Alone. *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*. Retrieved May 28, 2021, from <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.465.1338&rep=rep1&type=pdf>
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media Inc.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993–1022. Retrieved March 12, 2021, from <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>
- Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, 57(2), Article 7, 1–30. <https://doi.org/10.1145/1667053.1667056>

- Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM*, 55(4), 77–84.  
<https://doi.org/10.1145/2133806.2133826>
- Boutsidis, C., & Gallopoulos, E. (2008). SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4), 1350–1362.  
<https://doi.org/10.1016/j.patcog.2007.09.010>
- Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.
- Cambria, E., Havasi, C., & Hussain, A. (2012). SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis. *FLAIRS Conference*.
- Chien, J-T., Lee, C-H., & Tan, Z-H. (2018). Latent Dirichlet mixture model. *Neurocomputing*, 278, 12–22.  
<https://doi.org/10.1016/j.neucom.2017.08.029>
- Contreras-Piña, C., & Ríos, S. A. (2016). An empirical comparison of latent semantic models for applications in industry. *Neurocomputing*, 179, 176–185.  
<https://doi.org/10.1016/j.neucom.2015.11.080>
- Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6), 391–407.  
[https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6<391::AID-ASI1>3.0.CO;2-9](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9)
- Dumais, S. T. (2005). Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1), 188–230. <https://doi.org/10.1002/aris.1440380105>
- Elbagir, S., & Yang, J. (2019). Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment. *Proceedings of the International MultiConference of Engineers and Computer Scientists 2019 (IMECS 2019)*. Retrieved March 21, 2021, from [http://www.iaeng.org/publication/IMECS2019/IMECS2019\\_pp12-16.pdf](http://www.iaeng.org/publication/IMECS2019/IMECS2019_pp12-16.pdf)
- Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion mining. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, 417–422. Retrieved April 08, 2021, from [http://www.lrec-conf.org/proceedings/lrec2006/pdf/384\\_pdf.pdf](http://www.lrec-conf.org/proceedings/lrec2006/pdf/384_pdf.pdf)
- Fellbaum, C. A. (1998). Semantic Network of English: The Mother of All WordNets. *Computers and the Humanities*, 32, 209–220. <https://doi.org/10.1023/A:1001181927857>
- Gao, S., Tang, O., Wang, H., & Yin, P. (2018). Identifying competitors through comparative relation mining of online reviews in the restaurant industry. *International Journal of Hospitality Management*, 71, 19–32. <https://doi.org/10.1016/j.ijhm.2017.09.004>

- Gémar, G., & Jiménez-Quintero, JA. (2015). Text mining social media for competitive analysis. *Tourism & Management Studies*, 11(1), 84–90. Retrieved March 03, 2021 from <https://www.redalyc.org/pdf/3887/388743883010.pdf>
- Greene, D., O’Callaghan, D., & Cunningham, P. (2014). How Many Topics? Stability Analysis for Topic Models. In Calders, T., Esposito, F., Hüllermeier, E., & Meo, R. (eds), *Machine Learning and Knowledge Discovery in Databases*. ECML PKDD 2014. Lecture Notes in Computer Science, vol 8724. Springer, Berlin, Heidelberg. [https://doi.org/10.1007/978-3-662-44848-9\\_32](https://doi.org/10.1007/978-3-662-44848-9_32)
- Gu, X., & Kim, S. (2015). What Parts of Your Apps are Loved by Users? *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 760–770. <https://doi.org/10.1109/ASE.2015.57>
- Guo, Y., Barnes, S. J., & Jia, Q. (2017). Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation. *Tourism Management*, 59, 467–483. <https://doi.org/10.1016/j.tourman.2016.09.009>
- Guzman, E., & Maalej, W. (2014). How Do Users Like This Feature? A Fine Grained Sentiment Analysis of App Reviews. *2014 IEEE 22nd International Requirements Engineering Conference (RE)*, 153–162. <https://doi.org/10.1109/RE.2014.6912257>
- He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, 33(3), 464–472. <https://doi.org/10.1016/j.ijinfomgt.2013.01.001>
- He, W., Tian, X., Chen, Y., & Chong, D. (2016). Actionable Social Media Competitive Analytics For Understanding Customer Experiences. *Journal of Computer Information Systems*, 56(2), 145–155. <https://doi.org/10.1080/08874417.2016.1117377>
- Hofmann, T. (1999). Probabilistic latent semantic analysis. *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence (UAI1999)*, 289–296. Retrieved March 05, 2021, from <http://arxiv.org/abs/1301.6705>
- Hong, L., & Davison, B. D. (2010). Empirical study of topic modeling in Twitter. *Proceedings of the First Workshop on Social Media Analytics*, 80–88. <https://doi.org/10.1145/1964858.1964870>
- Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '04)*, 168–177. <https://doi.org/10.1145/1014052.1014073>
- Hu, F., & Trivedi, R. H. (2020). Mapping hotel brand positioning and competitive landscapes by text-mining user-generated content. *International Journal of Hospitality Management*, 84. <https://doi.org/10.1016/j.ijhm.2019.102317>



- Huebner, J., Frey, R. M., Ammendola, C., Fleisch, E., & Ilic, A. (2018). What People Like in Mobile Finance Apps: An Analysis of User Reviews. *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia*, 293–304. <https://doi.org/10.1145/3282894.3282895>
- Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1). Retrieved March 10, 2021, from <https://ojs.aaai.org/index.php/ICWSM/article/view/14550>
- Hyperreality. (n.d.). *British spellings*. Retrieved June 17, 2021, from [https://github.com/hyperreality/American-British-English-Translator/blob/master/data/british\\_spellings.json](https://github.com/hyperreality/American-British-English-Translator/blob/master/data/british_spellings.json)
- Iacob, C., & Harrison, R. (2013). Retrieving and analyzing mobile apps feature requests from online reviews. *2013 10th Working Conference on Mining Software Repositories (MSR)*, 41–44. <https://doi.org/10.1109/MSR.2013.6624001>
- Johann, T., Stanik, C., Alizadeh B., A. M., & Maalej, W. (2017). SAFE: A Simple Approach for Feature Extraction from App Descriptions and App Reviews. *2017 IEEE 25th International Requirements Engineering Conference (RE)*, 21–30. <https://doi.org/10.1109/RE.2017.71>
- Karpovich, S. (9 November 2013). *Probabilistic latent semantic analysis — PLSA*. Wikimedia Commons. [https://commons.wikimedia.org/wiki/File:Вероятностный\\_латентно-семантический\\_анализ.png](https://commons.wikimedia.org/wiki/File:Вероятностный_латентно-семантический_анализ.png)
- Kelly, Ryan. (2011). *PyEnchant*. <http://pyenchant.github.io/pyenchant/>. Accessed June 17, 2021.
- Kim, Y., & Jeong, S. R. (2018). Competitive intelligence in Korean Ramen Market using Text Mining and Sentiment Analysis. *Journal of Internet Computing and Services*, 19(1), 155–166. <https://doi.org/10.7472/JKSII.2018.19.1.155>
- Kooten, P. V. (n.d.). *Contractions*. <https://github.com/kootenpv/contractions>. Accessed June 17, 2021.
- Lafferty, J. D., McCallum, A., & Pereira, F.C.N. (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. *Proceedings of the Eighteenth International Conference on Machine Learning*, 282–289. Retrieved August 9, 2021, from [https://repository.upenn.edu/cgi/viewcontent.cgi?article=1162&context=cis\\_papers](https://repository.upenn.edu/cgi/viewcontent.cgi?article=1162&context=cis_papers)
- Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755), 788–791. <https://doi.org/10.1038/44565>
- Lee, JH., Park, S., Ahn, CM., & Kim, D. (2009). Automatic generic document summarization based on non-negative matrix factorization. *Information Processing & Management*, 45(1), 20–34. <https://doi.org/10.1016/j.ipm.2008.06.002>


- Li, Y., Jia, B., Guo, Y., & Chen, X. (2017). Mining User Reviews for Mobile App Comparisons. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3), 1–15. <https://doi.org/10.1145/3130935>
- Liu, B. (2012). Sentiment Analysis and Opinion Mining. *Synthesis Lectures on Human Language Technologies*, 5(1), 1–167. <https://doi.org/10.2200/s00416ed1v01y201204hlt016>
- Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5, 1608. <https://doi.org/10.1186/s40064-016-3252-8>
- Liu, Y., Jiang, C., & Zhao, H. (2019). Assessing product competitive advantages from the perspective of customers by mining user-generated content on social media. *Decision Support Systems*, 123. <https://doi.org/10.1016/j.dss.2019.113079>
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60. <https://doi.org/10.3115/v1/P14-5010>
- Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4), 1093–1113. <https://doi.org/10.1016/j.asej.2014.04.011>
- Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), 39–41. <https://doi.org/10.1145/219717.219748>
- Nielsen, F. Å. (2011, March 15). *A new ANEW: Evaluation of a word list for sentiment analysis in microblogs*. Retrieved March 15, 2021, from <https://arxiv.org/abs/1103.2903>
- Norman, G. J., Norris, C. J., Gollan, J., Ito, T. A., Hawkey, L. C., Larsen, J. T., Cacioppo, J. T., & Berntson, G. G. (2011). Current Emotion Research in Psychophysiology: The Neurobiology of Evaluative Bivalence. *Emotion Review*, 3(3), 349–359. <https://doi.org/10.1177/1754073911402403>
- O’Callaghan, D., Greene, D., Carthy, J., & Cunningham, P. (2015). An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13), 5645–5657. <https://doi.org/10.1016/j.eswa.2015.02.055>
- Osgood, C., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois.
- Ouyang, Y., Guo, B., Lu, X., Han, Q., Guo, T., & Yu, Z. (2019). CompetitiveBike: Competitive Analysis and Popularity Prediction of Bike-Sharing Apps Using Multi-Source Data. *IEEE Transactions on Mobile Computing*, 18(8), 1760–1773. <https://doi.org/10.1109/TMC.2018.2868933>
- Park, D. H., Liu, M., Zhai, C., & Wang, H. (2015). Leveraging User Reviews to Improve Accuracy for Mobile App Retrieval. *SIGIR '15: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 533–542. <https://doi.org/10.1145/2766462.2767759>

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12, 2825–2830.
- Pennebaker, J., Mehl, M., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1), 547–577.  
<https://doi.org/10.1146/annurev.psych.54.101601.145041>
- Porter, M. E. (1980). *Competitive Strategy: Techniques for Analyzing Industries and Competitors*. New York: Free Press.
- Qwertys. (21 October 2013). *Illustration of approximate non-negative matrix factorization (NMF)*, Wikimedia Commons. <https://commons.wikimedia.org/wiki/File:NMF.png>
- Santorini, B. (July 1990). *Part-of-Speech Tagging Guidelines for the Penn Treebank Project (3rd Revision)*. Retrieved May 08, 2021, from  
[https://repository.upenn.edu/cgi/viewcontent.cgi?article=1603&context=cis\\_reports](https://repository.upenn.edu/cgi/viewcontent.cgi?article=1603&context=cis_reports)
- Shah, F. A., Sirts, K., & Pfahl, D. (2019). Using app reviews for competitive analysis: tool support. *Proceedings of the 3rd ACM SIGSOFT International Workshop on App Market Analytics (WAMA 2019)*, 40–46. <https://doi.org/10.1145/3340496.3342756>
- Slxu.public. (29 September 2009). *Plate notation of the Smoothed LDA Model*. Wikimedia Commons. [https://commons.wikimedia.org/wiki/File:Smoothed\\_LDA.png](https://commons.wikimedia.org/wiki/File:Smoothed_LDA.png)
- Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The general inquirer: a computer approach to content analysis*. MIT Press, Cambridge.
- Su, Y., Wang, Y., & Yang, W. (2019). Mining and Comparing User Reviews across Similar Mobile Apps. *2019 15th International Conference on Mobile Ad-hoc and Sensor Networks (MSN)*, 338–342.  
<https://doi.org/10.1109/MSN48538.2019.00070>
- Suprayogi, E., Budi, I., & Mahendra, R. (2018). Information Extraction for Mobile Application User Review. *2018 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 343–348. <https://doi.org/10.1109/ICACSIS.2018.8618164>
- Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-Based Methods for Sentiment Analysis. *Computational Linguistics*, 37(2), 267–307. [https://doi.org/10.1162/coli\\_a\\_00049](https://doi.org/10.1162/coli_a_00049)
- Thelwall, M. (2013). *Heart and Soul: Sentiment Strength Detection in the Social Web with SentiStrength*. Retrieved March 28, 2021, from  
<http://sentistrength.wlv.ac.uk/documentation/SentiStrengthChapter.pdf>

- Truică, C. O., Velcin, J., & Boicea, A. (2015). Automatic Language Identification for Romance Languages Using Stop Words and Diacritics. *17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, 243–246. <https://doi.org/10.1109/SYNASC.2015.45>
- Truică, C., Rădulescu, F., & Boicea, A. (2016). Comparing Different Term Weighting Schemas for Topic Modeling. *2016 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, 307–310. <https://doi.org/10.1109/SYNASC.2016.055>
- Vu, P. M., Nguyen, T. T., Pham, H. V., & Nguyen, T. T. (2015). Mining User Opinions in Mobile App Reviews: A Keyword-Based Approach (T). *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 749–759. <https://doi.org/10.1109/ASE.2015.85>
- Wang, W., Wang, H., & Song, Y. (2016). Ranking product aspects through sentiment analysis of online reviews. *Journal of Experimental & Theoretical Artificial Intelligence*, 29(2), 227–246. <https://doi.org/10.1080/0952813X.2015.1132270>
- Wang, W., Feng, Y., & Dai, W. (2018). Topic analysis of online reviews for two competitive products using latent dirichlet allocation. *Electronic Commerce Research and Applications*, 29, 142–156, <https://doi.org/10.1016/j.elerap.2018.04.003>
- Whitelaw, C., Garg, N., & Argamon, S. (2005). Using appraisal groups for sentiment analysis. *Proceedings of the 14th ACM International Conference on Information and Knowledge Management - CIKM '05*. <https://doi.org/10.1145/1099554.1099714>
- Wilson, T., Wiebe, J., & Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 347–354. <https://doi.org/10.3115/1220575.1220619>
- Xu, K., Liao, S. S., Li, J., & Song, Y. (2011). Mining comparative opinions from customer reviews for Competitive Intelligence. *Decision Support Systems*, 50(4), 743–754. <https://doi.org/10.1016/j.dss.2010.08.021>

## Appendix A

Sampled reviews with wrong topics assigned by the NMF topic model

Review	Topic
It's still tell you that you got mail when you don't	10 - app/account deletion   download   account/login
My game is frozen katy60 help	2 - (group) chat   add feature   privacy
I got my statues saying showing that I'm online when I had it change to offline but still shows me being active online I don't like that I'm appearing offline for a reason to not be contact 	10 - app/account deletion   download   account/login
I can't understand is that a censorship or just stop responding period???	6 - app/feature working issue   apple watch
The past two weeks it's been kicking me out and taking me right back into my home screen on my phone Facebook has not responded to my report of this happening I'm missing my messages!	10 - app/account deletion   download   account/login
It's now going to be my third tome installing this app and it keeps telling me that I don't have Internet connection when I want to login please help as to what I should do ?	4 - app crash
Yeah very disgusted Messenger right now wouldn't let me into my Messenger so I had to uninstall it now I can't reinstall it keeps asking for payment information which I don't no what it has to do with it	4 - app crash
No longer have edit button option to update profile pic in messenger. When I search online for help, suggestions indicate to go through FB app. I do not use FB therefore do not have that app.	8 - status   search contact
Well I tried updating my messenger but it's been an hour and my internet is awesome I just got it fixed and it's still loading idk if it's just my phone or what but u don't think it is just saying	10 - app/account deletion   download   account/login
Why does Facecrook need another app on your phone or anything. So Mark can get paid. That's why ., oh and he can collect information about you.	10 - app/account deletion   download   account/login
Hi I logged into messenger with my Instagram account and I can't get logged into it. Plz help me. You guys should have told people who use messenger with a Instagram account and tell them we can't use the Instagram login account anymore or at least you guys should give me my account or something	7 - social/functional usage
All facebook messenger user Please be aware Armenian employers of Facebook used facebook for ban Azerbaijani users. They are taking advantage of working for Facebook to make censorship. Its intolerant and shoud be stopped	6 - app/feature working issue   apple watch
Since whatever update came through a few weeks ago, Messenger will no longer let me open links. Period. Doesn't matter if they're external or link back to FB. One star until you guys fix this crap. Multi-billion dollar company, and your app is trash.	9 - general problems to fix

## Appendix A (Continued)

I have used messenger for years, and out of no where I no longer receive notifications. It's infuriating as this is my main messaging app. I have tried everything from toggling settings to resetting my	6 - app/feature working issue   apple watch
free Palestine destroy Israel. free Palestine destroy Israel. free Palestine destroy Israel.	2 - (group) chat   add feature   privacy
Very bad Bad because of his racist policy against the Palestinians	2 - (group) chat   add feature   privacy
Facebook support ethnic cleansing against minorities worldwide by banning people, shutting, deleting, limiting audience for posts about oppressed people in Palestine and other areas in the world.	2 - (group) chat   add feature   privacy
#GazaUnderAttack #PalestineUnderAttack #Save_Sheikh_Jarrah	2 - (group) chat   add feature   privacy
Because this is facebook's then I give 1 star and also free Palestine	2 - (group) chat   add feature   privacy
The developers pretend to promote privacy, but sell sensitive data about your calls and texts to corporations. This is not shocking, as the app is owned by facebook.	5 - (video) call   connection quality
I am no longer getting notifications and sound when I receive text. I've reset to factory on phone and whatsapp and still nothing.	10 - app/account deletion   download   account/login
Merey phone mein chalna hai to warna achi tarha chal nahi to esa sabaq sikhaonga k munh dikhane k laik nai rahoge. Ye meri privacy hai aur meri privacy mein koi bhi entress nhi karna. Hukaayyy!!	2 - (group) chat   add feature   privacy
Terrible application. It sends your information to companies you don't want or to have.	11 - photo/video/link/message sending
WhatsApp says it is end to end encrypted and no one can read or hear not even WhatsApp. Well not true. I had someone hack my WhatsApp and get ahold of all my messages. So sad that you are advertising this. It's not true people clear your messages .	0 - message/notification receiving   read receipt
I already getting scammed and harassed on it	10 - app/account deletion   download   account/login
#القدس_تنتفض #التطبيع_خيانه #حي_الشيخ_جراح #لا_لتهود_القدس #أنقذوا_حي_الشيخ_جراح #لن_نرحل #انقذوا_حي_الشيخ_جراح #savesheikhjarrah #ŞeyhJarrahmahallesinkurtarın #SalvailquartierediSheikhJarrah #RettedasViertelSheikhJarrah #sauvezlequartierdesheikhjarrah	10 - app/account deletion   download   account/login
#gaza #save_palestine	2 - (group) chat   add feature   privacy
tracking my phone	10 - app/account deletion   download   account/login
None compare	2 - (group) chat   add feature   privacy
Awful app. I can't deactivate it. And they don't want you too. Stay away from this app. I tried to deactivate it. Impossible. Impossible to work, too.	6 - app/feature working issue   apple watch

## Appendix B

Sampled reviews with inconsistent labels of sentiment polarity on weighted sentiment scores

Review	autoAssigned_label	Manual_label	Remark
Trying to cancel a irder that was ordired incorrectly, please help me cancel the order I have tried everything. Order is from Linda vickers in amount of \$29 . Please help	Positive	Negative	
Recently my messages won't go through, it'll just say sending but will take roughly 10-20 mins to finally send. My internet works perfectly fine and so does my data so I think the app may need another fix/update asap!	Positive	Negative	
Great app. Since the update the Bluetooth no longer works with my tesla speakers when video is on (fine- - video probably shouldnt be on anyways). Fix it please	Positive	Negative	Positive at first, then shift to app problems
I need this app for calling	Negative	Positive	
Support animated stickers	Negative	Positive	Advice
This is an app that doesnt disappoint. love it!	Negative	Positive	
Very helpful for communication across the world or just iPhone to android when relying on WiFi in dead zones	Negative	Positive	
it connect s me with my favorite people. thank you. Lolinche	Negative	Positive	
I'll start by saying the app is pretty awesome. I only have a couple things that bug me persistently. The first is a lack of timestamps. The recent messages read "10 hours ago" for example. I'd like to see an actual time so I don't have to do math for the messages sent / received within 24 hours. Also, messages show as sent and received on my end, but the recipient actually doesn't get them for hours. And sometimes the same thing happens with me on the receiving end.	Positive	Negative	Positive at first, then shift to app problems
Love the app. Use it daily. Access to photos not available in IOS 14. Goes to "recent" folder to find photo album. No pictures in this folder.	Positive	Negative	Positive at first, then shift to app problems
I switched from WeChat and WhatsApp to Signal.	Negative	Positive	
Could be better if you guys give options of status or history. Also the profile picture option needs to be improved.	Negative	Positive	Advice

