Sentiment Analysis of Mobile Apps Using BERT

Wajhee Ullah, Zheying Zhang, and Kostas Stefanidis

Tampere University, Finland
{wajhee.ullah,zheying.zhang,konstantinos.stefanidis}@tuni.fi

Abstract. In this paper, we focus on identifying issues in mobile app updates that adversely impact the opinion in user reviews by analyzing the sentiment of the reviews. We use sentiment analysis using BERT to evaluate the performance of mobile apps and the sentiment distribution of reviews for identifying the cause of sentiment shifts. Using our method, developers can correctly locate the period of specific sentiment and review the sentences and keywords used in reviews to identify the problems and complaints in recent updates. An increase in negative sentiments after any major update can help identify the exact issue causing the problem. Our experimental analysis shows the effectiveness of the proposed method in recognizing issues and identifying any potential problematic updates.

Keywords: Sentiment Classification, BERT, Mobile Apps

1 Introduction

The rapid development of distributed computing has enabled the analysis of vast amounts of data and the prediction of customer preferences and demands. Understanding customers' emotional inclinations towards the application that they use daily has become increasingly important [17,10,9,11]. User reviews and ratings are critical factors in app selection, with studies indicating that users typically download apps based on these factors [7]. The reviews also help developers gain insight into user sentiment about applications and help to make decisions in rolling out updates to address issues or introduce new features.Sentiment analysis, a technique that extracts opinions from text, is a very powerful tool for finding out the emotions hidden behind the review and feedback.

In this work, we focus on identifying issues in mobile app updates that adversely impact the opinion in user reviews by analyzing the sentiment of user reviews. We classify the app reviews into three categories: positive, negative, and neutral, and use this distinction to identify the problematic mobile app updates based on the number of negative reviews in a certain period. This approach allows developers to be aware of user opinions on app updates and guides them proactively to address the most important issues early.

To perform sentiment analysis, we utilize the BERT model, namely the Bidirectional Encoder Representation from Transformers [4]. BERT is a deep learning model where weights between elements are dynamically determined depending

on their relationship. It uses an encoder to read text input and a decoder to provide predictions. The transformer, which is the attention mechanism that learns contextual relationships between words in a text, is integral to the model. The encoder processes a series of tokens to produce an output transformed into vectors and used for sentiment analysis. The paper's approach is essential because negative reviews can significantly impact the success of a mobile app. Therefore, it is crucial to identify any issues that lead to negative sentiment and address them promptly. The proposed method is effective in recognizing issues and identifying any potentially problematic updates. By leveraging the approach, developers can improve their mobile apps, enhance user experience, and increase user satisfaction.

2 Related Work

Sentiment analysis research has been extensively reported in the literature. Many studies applied lexicon-based approaches to extracting words that express positive or negative feelings in the text and analyzing the overall opinion by aggregating the sentiment score of these words. This approach uses a given lexical database or corpus-based lexicons tailored to specific domains, and each word is labeled as positive, negative, or neutral sentiments along with polarity. For example, [3] retrieved adjectives in the Amazon and CNET datasets and analyze the sentiment of reviews based on the positive or negative polarity of adjectives, using WordNet[1]. [5] investigated different techniques for calculating prior polarity scores based on SentiWordNet. [12] presents a Lexicon-based approach that considers positive, comparative, and superlative comparisons. This dictionary-based technique matches words inside phrases to determine their polarity by matching emotional lexis with both positive and negative terms. It uses a combination of lexicon heuristics and a pre-trained model to analyze text and provide sentiment scores. It also takes into account the intensity of sentiment. Vader [8] also provides a compound score which is a normalized, weighted composite score. This score is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized. When information is scarce, this lexicon could provide useful polarity alternatives to machine learning methodologies, and it is imperative to consider other workable possibilities.

In addition to the traditional lexicon-based methods, machine learning approaches are widely used for studies on sentiment analysis. [15] investigated pre-processing techniques for sentiment analysis of Twitter datasets using four machine learning algorithms. [2] uses *Bidirectional Encoder Representations from Transformers* (BERT) to analyze emotions. Four machine learning algorithms are utilized to compare with the BERT. Overall, experiments have shown that BERT surpasses machine learning approaches with socially constructed baselines for emotional analysis. For TextCNN, accuracy slightly increases than other machine learning techniques. Although the differences are not large, SVM gives better results when compared to Naive Bayes and *k-Nearest Neighbors algorithm*

[1] https://wordnet.princeton.edu/

(k-NN) in terms of relative performance. [6] conducted a comparative analysis of four commonly used algorithms, and they are Naive Bayes, Max Entropy, Boosted Trees, and Random Forest Algorithms. This study reported that, although requiring a lot of training time, the Random Forest classifier has good accuracy and performance, is easy to comprehend, and consistently produces incremental results over time. The NB classifier uses less memory and takes less time to train. Max Entropy is a worthy alternative if less training time is the highest priority. On average, when considering different aspects, the best-performing classifier is the boosted Tress. Additionally, [14] proposed a hybrid approach by combining rule-based classification, supervised learning, and machine learning into a new combined method, and tested it on movie reviews, product reviews, and MySpace comments.

The feature extraction methods employed in the emotional analysis had undergone a fair amount of adjustment [13]. The Stanford parser is used to parse movie reviews. Four feature extraction strategies are applied for feature extraction across many corpora. This approach is called the *Intrinsic Extrinsic Domain Relevance Approach* (IEDR), it is not industry-specific in its design, and when compared to other methodologies utilized for sentiment classification and analysis, it suggests feature extraction for performance improvements. A proportional examination of emotional location using *Support Vector Machine* (SVM) and Naive Bayes techniques was conducted in [1]. The best accuracy is obtained when utilizing Naive Bayes with the synthetic word method and linear SVM. An evaluation study on movies shows that dramatic film genres distinguish out for their greater authenticity when compared to other film genres. Using a graph, the authors also display the polarity of certain words.

3 The BERT Model

BERT is an open-source machine learning framework for natural language processing. It targets grasping the meaning of uncertain words in a phrase or sentence by building meaning from adjacent information. Originally, it was trained using 2500M words from English Wikipedia and 800M words from the BooksCorpus Dataset. BERT can operate across both directions while simultaneously reading, i.e., can understand the text from both left to right or right to left. Using this bidirectional capacity, BERT is pre-trained on two different but related natural language processing tasks. In turn, *Masked Language Model* (MLM) conceals some percentage of the input tokens at random to predict the original vocabulary of the concealed word based only on its context. The model enables pre-trained deep bidirectional representations. The *Next Sentence Prediction* (NSP) task trains a model that understands sentence relationships.

The main purpose of almost any natural language processing method is to understand human language. For achieving this, traditional models are usually trained using a large pool of exact data. BERT, on the other hand, is trained with just an unstructured simple text collection. Despite its use in real-world

scenarios, it continues to acquire knowledge unattended from sequences of words. The pre-processing serves as any knowledge ground.

BERT's model architecture is a multi-layer bidirectional Transformer encoder. This transformer is the model component that enables BERT to detect background and uncertainty. It is achieved by the transformer by assessing each phrase and comparing it to every other part of a phrase. Examining the related letters assists BERT in grasping the full background of the word, helping it to better comprehend the searcher's intent.

In contrast to the traditional word embedding strategy, older systems might convert each singular phrase into such a directional quantity that conveyed nothing but one fragment regarding the definition of words. Such embedding models demand a large amount of labeled data. While they succeed at several broad natural language processing tasks, they fail to demonstrate effectiveness, due to the fact that every word is in one way or another linked to a definition.

BERT takes into account the use of MLM to prevent the required word from *seeing itself* and acquiring a limited interpretation irrespective of its surroundings. As a result, BERT only considers camouflaged text by looking at the preface. In this framework, texts are based on their conditions rather than by being identified upon some static approach.

This note is of great importance since words frequently alter meaning as a phrase progresses. Each new word adds to the overall significance of the term targeted by natural language processing mechanisms. The more words in each piece of writing, the less evident the term in focus becomes. By reading two-way communications, taking into consideration the impact of all other words in the sentence on the target word, and minimizing the movement that causes words to have a specific connotation as a sentence progresses, BERT gives enhanced meaning.

Prediction of next phrase. During the BERT training process, the system is given a series of inputs and is asked to do some calculations and guess whether the next input of the series is the next word of the given data. Half of the values are in the form of pairs in which the next value is the next sentence of the original input value. The remaining 50% of the total are arbitrary words from the corpus. It is expected that the randomized text is distinct from the first sentence. To assist the model in distinguishing between the two phrases during training, the data is handled as follows before reaching the model:

- A token appears at the start of the first sentence, and a token appears just at the end of every text. Token embeddings are a way to represent words or other discrete units of text as dense vectors of real numbers.
- Each token receives a sentence embedding. Sentences with a range of 2, extracted features are essentially equivalent to token word embeddings.
- A positional embedding is assigned to each token to signify its location in the paragraph. Transformer describes the idea and execution of directional embedding.

To predict if the second sentence is connected to the first, the following procedure is carried out:

- The Transformer model goes through the whole input sequence.
- The [CLS] token output is transformed into a 2*1 shaped vector using a basic classification layer (learned matrices of weights and biases).
- The probability of something being the next sequence is predicted by Soft-Max.

In the BERT model, MLM and NSP are trained concurrently with the goal of minimizing the combined loss function of the two approaches. The data for this classification job must be split into two parts: test sample and training batch. The learning is required to train the classifier, and the second sample is used to evaluate the classifier's performance.

The sentiment classifier is built on top of the BERT model. we took advantage of BERT's ability to encode and understand the contextual relationships between words in a sentence. To adapt BERT for sentiment classification, we fine-tuned the pre-trained model on a labeled dataset of reviews and their corresponding sentiment labels. During the training process, the weights of the BERT model are updated on the basis of input data for sentiment classification which allow the model to learn the sentiments.

4 Experimental Setting

In this section, we study how to identify the abnormal days and the mobile app updates that adversely affect the sentiments of user reviews.

Dataset. The dataset contains user reviews from five social media mobile apps, collected using google-play-scraper[2]. The reviews for each app are from September 1, 2016, to August 31, 2017. Altogether, the dataset contains 202,870 reviews for IMO, 122,622 reviews for Hangouts, 1,654,360 reviews for Messenger, 153,128 reviews for Skype, and 1,660,145 reviews for Whatsapp.

Pre-processing. During the pre-processing phase, non-English reviews are filtered out. We eliminate non-English reviews with Langdetect[3], a handy language detection program for Python.

Mobile app reviews are usually short compared to the reviews on other platforms but still, the multiple sentences in a single review could affect the sentiment and mean [3]. The next step in pre-processing is to separate the multiple sentences in a single review into individual sentences. For this purpose, sentence tokenizers are used to split the text or whole reviews into individual sentences. We have used the *Natural Language Toolkit* (NLTK) python package to split the reviews into sentences. NLTK is a leading platform for building Python programs to work with human language data.

Table 1 shows the statistics for the datasets. Imo and Hangout contain fewer reviews compared to Messenger and WhatsApp which have over 1M reviews. Reviews are classified into positive, negative, or neutral based on their rating score, where a review containing a rating score of 4 or 5 is classified as Positive,

[2] https://github.com/JoMingyu/google-play-scraper
[3] https://pypi.org/project/langdetect/

Table 1. Apps Reviews Statistics

App Name	Reviews	English Reviews	Sentences	Updates
Imo	202,870	86,194	100,838	84
Hangouts	122,622	68,535	101,704	43
Messenger	1,654,360	886,643	1,185,368	105
Skype	153,128	105,875	189,995	76
Whatsapp	1,660,145	851,662	1,098,583	49

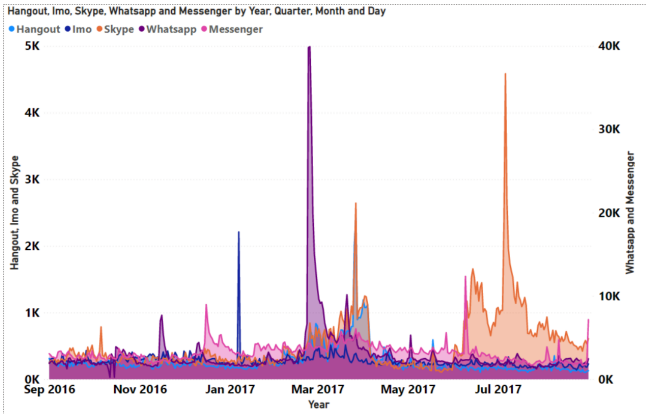

Fig. 1. Frequency of Apps Reviews over Time

a review containing a rating score of 3 is classified as Neutral, and a review with a rating score 1 or 2 is categorized as Negative. Together with the reviews, the number of updates of each mobile app in the given period is also shown in the table. Also, Figure 1 shows the frequency of all app reviews over the period of time. Finally, we classify the dataset into training and validation sets. The training set consists of 85% of the original dataset, while the validation set contains the remaining 15%.

Training. For training, the learning rate, the weight of the network with respect to the loss gradient descent, is 0.001. It determines how fast or slow we will move toward the optimal weights. The dropout for the training model is set at 0.5. The dropout layer prevents the overfitting of the model by ignoring some neurons during the training phase. We used Adam as our optimizer, CategoricalCrossentropy as our loss function, and SparseCategoricalAccuracy as our accuracy metric. The training model is trained for 10 epochs.

5 Evaluation

5.1 Identify Abnormal Days

In this section, we analyze the sentiment of the users' reviews of 5 apps, from 1 September 2016 to 31 August 2017. The abnormal distribution of some specific

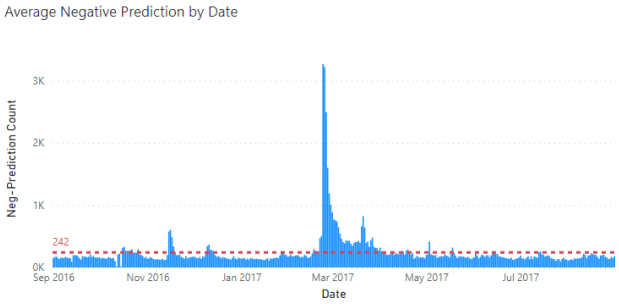

Fig. 2. Predicted Negative Trends by days – Whatsapp

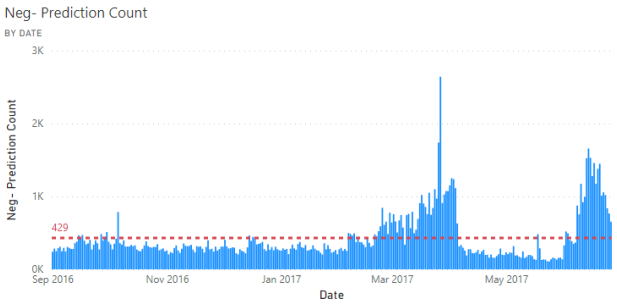

Fig. 3. Predicted Negative Trends by days – Skype

days helps to identify the days with most probably some problem in the app. The distribution of negative reviews of individual apps and analysis are shown below.

Whatsapp. Based on the obtained results, the count of negative value trended up and resulting in a 21.19% increase between Thursday, September 1, 2016, and Thursday, August 31, 2017. Also from the predicted values, the negative value started trending down on Wednesday, March 29, 2017, falling by 41.53% (130) in 5.07 months. It can also be concluded from Fig. 2 that the count of negative values jumped from 251 to 1,007 during its steepest incline between Monday, February 20, 2017, and Tuesday, February 28, 2017.

Skype. The trends obtained from the predicted values in Fig. 3 show that the average negative prediction trended up, resulting in a 217.84% increase between Thursday, September 1, 2016, and Sunday, June 25, 2017, while the average negative prediction started trending down on Saturday, June 10, 2017, falling by 34.70% (407) in 15 days. It is also evident from the results that the average negative prediction jumped from 368 to 755 during its steepest incline between Wednesday, June 7, 2017, and Friday, June 9, 2017.

Hangout. Fig. 4 shows the average negative prediction trended down, resulting in a 62.96% decrease between Thursday, September 1, 2016, and Thursday, Au-

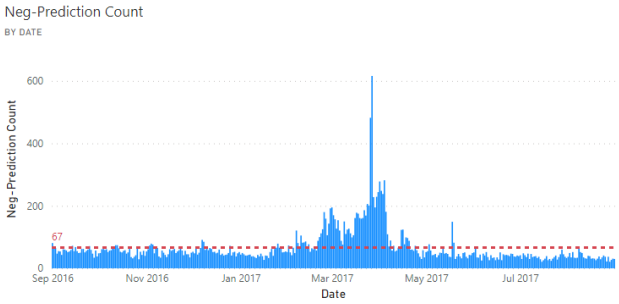

Fig. 4. Predicted Negative Trends by days – Hangout

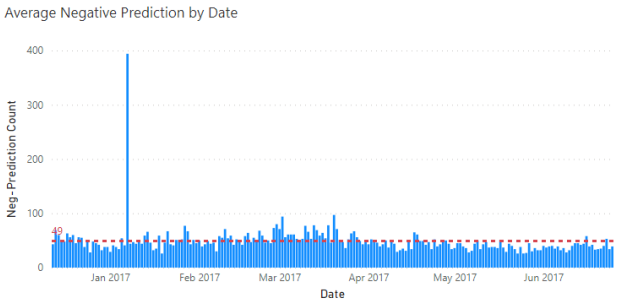

Fig. 5. Predicted Negative Trends by days – Imo

gust 31, 2017. It can also be concluded that average negative prediction started trending down on Sunday, April 23, 2017, falling by 46.43% (26) in 4.27 months while negative prediction gained during its steepest incline between Thursday, March 30, 2017, and Sunday, April 30, 2017.

Imo. Fig. 5 shows the average negative prediction trended down, resulting in a 16.00% decrease between Thursday, September 1, 2016, and Thursday, August 31 2017 also average negative prediction started trending down on Tuesday, January 10, 2017, falling by 4.55% in 7.70 months. It is also evident that the average negative prediction jumped from 38 to 47 during its steepest incline between Tuesday, January 3, 2017, and Monday, January 9, 2017.

Messenger Fig. 6 shows that the average negative prediction trended down, resulting in a 1.57% decrease between Monday, December 12, 2016, and Sunday, June 25 2017 also the average negative prediction started trending down on Monday, June 12, 2017, falling by 0.30% in 12 days. It is also shown from the results that the average negative prediction jumped from 977 to 1,066 during its steepest incline between Monday, June 5, 2017, and Sunday, June 11, 2017.

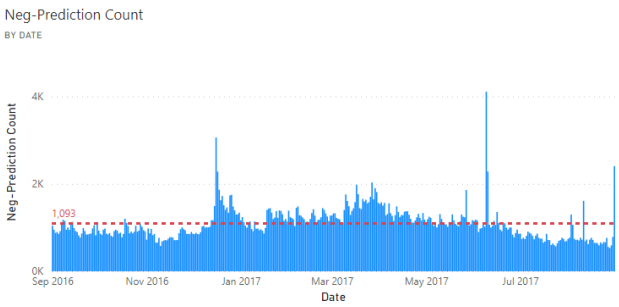

Fig. 6. Predicted Negative Trends by days – Messenger

5.2 Identify Abnormal Updates

The results obtained are used to identify the days with abnormal negative reviews. Further investigation of the negative reviews and the trends helps to identify the potential abnormal updates. The update history of the apps and their relationship with the negative review count could help to identify the adverse impact of updates and further exploring the reviews would also provide the reasons for the issue which could help the developers to identify the potential problems in app-updates and cause of application's low rating. We have manually analyzed the reviews about the days that have an abnormal amount of negative reviews predicted by the algorithm. Reviews on those specific days provide insight into potential issues in the apps which are further discussed below.

Whatsapp. Based on the results of analyzing the number of negative reviews on WhatsApp and the update history of the app, we found that the update on February 22, 2017 overlaps with the dates when the number of negative reviews increases the most. The update introduced the story feature that enables users to share status messages in the form of a GIF, video, or an image[4]. The results of sentiment analysis show the number of negative reviews jumped from 251 to 1007, which is the steepest incline in negative reviews between February 20, 2017, to February 28, 2017. This could be argued that the update may have been responsible for the sudden increase in negative reviews and be identified as a problematic update, which can be confirmed by looking at the example reviews in Table 2. The newly added story feature and the change to the location of the contacts tab are not appreciated by many users.

Imo. From the predicted results for the Imo application, we looked into the duration from January 03, 2017, to January 09, 2017, and investigate further the reasons that cause the increased negative reviews. The previous update from the negative spike update between January 03 2017 to January 09 2017 was

[4] https://www.businesstoday.in/technology/news/story/whatsapp-changes-everything-with-its-new-status-feature-71508-2017-02-21

Table 2. WhatsApp Negative reviews after new update

Date	Review	Rating
22/02/2017	"New update is so bad, deleted the photo or video story update, its not Instagram or line apps, as long as I know, whats app is only chatting apps"	1
22/02/2017	how to stop new status updates, I don't want to see people status"	1
23/02/2017	"I am looking forward to yhe new update where there is a separate status and story tab for both of them"	1

Table 3. Imo reviews after increased negative review count

Date	Review	Rating
07/01/2017	"Something happened to where I had to reinstall the app and after that I am not able to verify my number on it to reinstall"	2
07/01/2017	"My imo acount not working properly even I uninstall app and installed again not but it doesn't working and take a long time for verification code after that it shut down automatically"	2
08/01/2017	"Update versions is not good..."	2

on December 21, 2016 (9.8.000000004201 (1248)) so we can conclude that the previous update was not the cause for increased negative reviews. But analyzing the reviews of this duration it could be concluded that failure in some app functionality and abnormal closing of the app is the main issue users were facing in this duration.

Skype. According to the results of the skype application, in Early, June 2017 there is an abnormal no. of negative reviews from users. Review analysis from these dates could help us to identify the problem. Skype had an update at the start of June 2017[5] and the negative reviews trended upwards since the start of June 2017. The negative reviews mostly state the issue faced after the app update. Some users were worried about the new UI introduced in the application while some reviews are due to technical issues in the application like not being able to send a file etc.

Hangouts. From the results of the Hangouts application, we have a negative review increase at the end of February 2017 and the beginning of March 2017. Further study of the reviews helps us to identify the reasons. Hangouts had an update on February 27, 2017[6] and the negative reviews trended upwards after then and in the start of March 2017. The negative reviews mostly stated the issue about the forcefully close and crash of the application while using.

Messenger. According to the results of the Messenger application, in Dec 2016 there is a sudden increase in the number of negative reviews from users. Further review analysis from these dates could help us to identify the problem. Messenger

[5] https://www.greenbot.com/microsoft-completely-revamps-skype-new-ui-snapchat-like-stories/

[6] https://www.apkmirror.com/apk/google-inc/hangouts/hangouts-17-0-148298972-release/hangouts-17-0-148298972-17-android-apk-download/

Table 4. Skype reviews after increased negative review count

Date	Review	Rating
08/06/2017	"If these kinda updates continue then doom's day is not far away skype."	1
08/06/2017	"Cannot send file on new version ?!?!?!?!"	1
04/06/2017	"Terrible new UI treating me like I have reading difficulties, displaying everything very big and feels very sticky."	2

Table 5. Hangouts reviews after increased negative review count

Date	Review	Rating
03/03/2017	"Really wish Google had better quality control and a better vision for its products."	1
03/03/2017	"Its a great app and u should dowload it but plz fix the bugs it wasnt working when i was chating with my friend and we couldnt video call because it say on going call when we hung up"	1
02/03/2017	"Why y'all forcing me to update my shit I'm good with the old version"	1

has an update on 15 December 201 and 8 March 2017[7] and we see the negative reviews increased after that. The reviews mention a problem with the battery drain which could be an ongoing issue with the application but the problem with the photo uploading that takes much longer could be the issue that has been coming with the update.

6 Conclusion

This paper presents a sentiment analysis of reviews for five different mobile apps using BERT. Sentiment analysis is an effective method to evaluate the performance of applications, and the sentiment distribution of reviews is particularly useful for identifying the cause of sentiment shifts. Developers can use this method to correctly identify the period of specific sentiment and further review the sentences and keywords used in reviews to identify the problems and complaints in recent updates. An increase in negative sentiments after any major update can help identify the exact issue causing the problem. This method presented in the paper aims to help developers stay informed about any major issues that cause low ratings in user reviews and identify any potential update that adversely impacts the sentiment of reviews. The results demonstrate the effectiveness of the proposed method in recognizing issues and identifying any potential problematic updates.

Acknowledgement

This work is based on the Master Thesis "Sentiment Analysis of Mobile Apps Using BERT" [16], appearing at the digital library of the Tampere University. We thank the Tampere University for supporting this research.

[7] https://www.apkmirror.com/apk/facebook-2/messenger/messenger-100-0-0-29-61-release/

Table 6. Messenger reviews after increased negative review count

Date	Review	Rating
11/03/2017	"Would be fine if it wasn't draining my battery."	2
11/02/2017	"It takes extremely long to send an image which I never had a proboem in the past."	1
11/03/2017	"Should be fast and photo uploading new ways should be introduced .	2

References

1. Alnashwan, R., O'Riordan, A., Sorensen, H., Hoare, C.: Improving sentiment analysis through ensemble learning of meta-level features. In: KDWeb (2016)
2. Cao, Y., Sun, Z., Li, L., Mo, W.: A study of sentiment analysis algorithms for agricultural product reviews based on improved BERT model. Symmetry **14**(8), 1604 (2022)
3. Cui, H., Mittal, V.O., Datar, M.: Comparative experiments on sentiment classification for online product reviews. In: Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference (2006)
4. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (2019)
5. Guerini, M., Gatti, L., Turchi, M.: Sentiment analysis: How to derive prior polarities from sentiwordnet. In: EMNLP (2013)
6. Gupte, A., Joshi, S., Gadgul, P., Kadam, A.: Comparative study of classification algorithms used in sentiment analysis. IJCSIT **5**(5) (2014)
7. Harman, M., Jia, Y., Zhang, Y.: App store mining and analysis: MSR for app stores. In: MSR (2012)
8. Hutto, C.J., Gilbert, E.: VADER: A parsimonious rule-based model for sentiment analysis of social media text. In: ICWSM. The AAAI Press (2014)
9. Li, X., Zhang, B., Zhang, Z., Stefanidis, K.: A sentiment-statistical approach for identifying problematic mobile app updates based on user reviews. Inf. **11**(3), 152 (2020)
10. Li, X., Zhang, Z., Stefanidis, K.: Mobile app evolution analysis based on user reviews. In: SoMeT. Frontiers in Artificial Intelligence and Applications (2018)
11. Li, X., Zhang, Z., Stefanidis, K.: A data-driven approach for video game playability analysis based on players' reviews. Inf. **12**(3), 129 (2021)
12. Mandal, S., Gupta, S.: A novel dictionary-based classification algorithm for opinion mining. In: ICRCICN (2016)
13. Pasarate, S., Shedge, R.: Comparative study of feature extraction techniques used in sentiment analysis. In: ICICCS-INBUSH (2016)
14. Prabowo, R., Thelwall, M.: Sentiment analysis: A combined approach. J. Informetrics **3**(2) (2009)
15. Symeonidis, S., Effrosynidis, D., Arampatzis, A.: A comparative evaluation of preprocessing techniques and their interactions for twitter sentiment analysis. Expert Syst. Appl. **110** (2018)
16. Ullah, W.: Sentiment analysis of mobile apps using BERT. Master of Science Thesis, Faculty of Information Technology and Communication Sciences, Tampere University, Finland (January 2023)
17. Zhang, L., Hua, K., Wang, H., Qian, G., Zhang, L.: Sentiment analysis on reviews of mobile users. In: MobiSPC (2014)