2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService)

Sentiment Analysis of Chinese Product Reviews using Gated Recurrent Unit

Jun Sheng, Lee
Business Analytics Centre,
National University of Singapore
Logitech Europe S.A.
Singapore, Singapore
junsheng@u.nus.edu

Denis Zuba
Data Science & Advanced Analytics,
Logitech Europe S.A.
Lausanne, Switzerland
dzuba@logitech.com

Yan, Pang
Department of Analytics and Operations,
National University of Singapore
Singapore, Singapore
jamespang@nus.edu.sg

***Abstract*—Despite the explosive growth of Chinese e-commerce platforms in recent years, research focusing on the sentiment classification of Chinese documents pales in comparison to its western counterparts (English documents). This paper looks into the nascent area of Natural Language Processing (NLP) in the Sentiment Analysis of Chinese Text. The proposed Deep Learning method is the use of a sentence-based approach in the sentiment analysis of online reviews to gain more granularity and increased classification accuracy. Experimental results on a balanced (50:50), 2 class (positive, negative) test dataset of 1669 product reviews show an empirical accuracy of 87.66%, while results on an imbalanced (18:82) test dataset of 2519 product reviews show an accuracy of 87.9%, thus demonstrating the effectiveness and robustness of this proposed approach.**

Keywords—Deep Learning, Chinese Sentiment Analysis, Sentiment Analysis, Natural Language Processing, Text Analytics, Text Mining, Gated Recurrent Unit, Neural Networks

I. INTRODUCTION

Today, the world produces an astronomical amount of data. Every day, 18.7 billion text messages are sent out globally [1]. Facebook, the world's largest social networking platform with 2.2 billion active users, has 350 million new photos uploaded daily [2]. Google processes 40,000 search queries every second [3]. The International Data Corporation (IDC) forecasts that the amount of data generated worldwide would reach a staggering 44 zettabytes by 2020, 50 times the amount of data generated in 2010 [4]. All these numbers point to the same story: the rapid growth of data has taken the globe by storm, and this growth shows no signs of slowing down anytime soon.

Data generally can be classified into two main categories. The first being structured data, which consists of information with clearly defined data types. The second category of data refers to unstructured data, i.e. data not organized in a pre-defined manner, mainly composed of data of social media platforms, emails, videos, etc. According to estimates by experts, unstructured data would account for 93% of all data in the universe by 2022 [5].

One form of unstructured data that has increasingly garnered business attention is User-Generated Content (UGC), i.e. publicly available content created by users of a service/system. Examples of UGC include comments, reviews, recommendations on blogs, tweets, and photos etc. Today, consumers are rapidly tuning out traditional media (TV and radio), instead, spending more time online through their mobile devices. Statistics show that the average millennial today spends 5.4 hours/day looking at UGC [6].

With increasing availability of UGC, consumers are also increasingly turning to referrals and recommendations before making a purchase [7]. Nielsen Consumer Trust Index reports that up to 92% of consumers now trust organic UGC more than traditional advertising [8].

Overall, with the emergence of many information channels over the last decade, UGC has fundamentally changed the purchasing behaviors of consumers worldwide, and could very well be the new "oil" of the 21st century. If harnessed well, UGC can be used to unlock new sources of consumer insights, inject fresh business ideas, and fuel the next wave of productivity and economic growth [9].

This paper proposes a sentiment analysis approach based on deep learning technology to uncover affective states (positive/negative) in unstructured data. Specifically, it will explore the use of sentiment analysis on the nascent area of Chinese texts. With rapid advances in the area of machine learning, deep learning has emerged as a leading method used in analyzing unstructured data. One way of applying deep learning is through sentiment analysis.

Sentiment analysis is a typical text classification method that allows mining of text to identify and extract subjective information in the source material. Through sentiment analysis, companies stand to yield immense benefits in understanding the social sentiment of their brand, product or service.

The paper will be structured as follows. A comprehensive literature review will be outlined in section 2. In section 3, the proposed approach for sentence level sentiment analysis will be studied in detail. Following which, in section 4, the techniques used for evaluation are defined and results from the proposed method be shown. Section 5 discusses the value achieved by the paper. Finally, conclusions and areas for further work will round up the paper in section 6 and 7.

II. LITERATURE REVIEW

There are two methods used for sentiment analysis, Lexicon-based approach (linguistic method) and the Machine Learning approach (statistical/computing method).

A. Lexicon-based Approach

Lexicon-based approach is an unsupervised learning method that calculates the orientation of a document based on the semantic orientation of words in the document [10]. This method uses a dictionary of words annotated with a word's semantic orientation, aka. polarity (positive, negative, neutral) and sentiment strength. Lexicon-based approach involves the manual creation of dictionaries [11] [12].

978-1-7281-0059-3/19/$31.00 ©2019 IEEE
DOI 10.1109/BigDataService.2019.00030

Currently, most of the lexicon-based research focus on the semantic orientation of the adjectives in the document [13][14][15][16]. One key advantage of the Lexicon-based approach is that no labelling and training of data is required. Furthermore, it achieves relatively strong performance, with a well-developed, lexicon-based approach usually achieving a good accuracy between 70-80%.

In term of disadvantages, this method requires extensive linguistic resources which may not always be available, as is the case for sentiment analysis of Chinese text. Also, especially in the area of social media analytics, dictionaries are not yet developed to adapt to the linguistic and para-linguistic features of computer-mediated communication.

B. Machine Learning Approach

Machine Learning approach is a method that builds classifiers from labelled datasets of documents [17]. This approach makes use of labelled datasets and is a form of supervised learning method, whereby each instance in the dataset is a pair consisting of an input object (document) and its label (sentiment polarity).

Support Vector Machines of Linear Classifiers, Naïve Bayes and N-grams of Probabilistic Classifiers are some of the more popular methods used in text classification algorithms and are generally grouped as traditional machine learning approaches.

The main advantage of this method is that it does not require usage of a dictionary. Performance of such classifiers also achieve excellent accuracy (80-85%) in sentiment analysis [18][19][20][21].

A disadvantage is that these trained classifiers are generally domain specific and only achieves good performance if it is used within its field. Furthermore, it is reported that performance drops notably when applied to text from other domains. [22][23]. Another disadvantage would be the need to obtain labelled training data, but in comparison, this task is not as resource intensive as the obtaining a dictionary for lexicon based approaches.

In summary, the choice of the method used (Lexicon-based, Machine Learning) usually depends on several considerations, including the application, domain, language and resources available.

C. Sentiment Analysis using Deep Learning Techniques

In recent years, deep learning has garnered increasing attention within the industry and academic world for its high performance in areas, e.g. computer vision, speech/audio recognition and more importantly, Natural Language Processing (NLP). The performance of deep learning in NLP tasks has been shown to outperform traditional machine learning approaches for sentiment analysis. [24][25][26].

One critical aspect of Deep Learning is its ability to learn features from raw data automatically. Deep Learning effectively bypasses the need for manual feature engineering, in turn producing insightful discoveries and uncovering hidden patterns within the data.

Deep learning works by using a cascade of multiple layers of nonlinear processing units to extract features and perform a transformation on the data that is fed into it. At each successive layer, it uses the output from the previous layer as its input and transforms the input data into a more abstract and composite representation before feeding it to the next layer. At each layer, this process helps network learn new and different features of the data. This architecture allows for a powerful hierarchical and feature representation.

Fig. 1. Deep Learning Feature Hierarchy

Fig. 1 depicts how the feature hierarchy develops from the first layers on the left to its final layer on the right learned by a deep learning model for the image classification of faces [27]. As illustrated, the image features grow in complexity, starting from simple features like edges, then to more complex features like parts of a face, e.g. the nose, eyes, ears, and finally into faces in the later layers. This concept can be applied to text as well.

The main advantage of deep learning is that it yields the highest accuracies amongst all methods and the removal of manual feature engineering as a pre-processing step. This may explain deep learning's outperformance and popularity vis-à-vis traditional machine learning approaches which are feature-based.

However, one disadvantage and a vital characteristic of this approach is that large amounts of labelled data are required for training in order to allow the model to achieve high accuracies [28]. Another disadvantage is training time. Training a deep learning model can be a challenging and expensive task to undertake. Nonetheless, significant development has been made in Graphics Processing Units (GPUs), shortening training times and speeding up calculations.

In the midst of advances in the field of Sentiment Analysis on English Text, deep learning techniques have emerged as the gold standard for its outperformance (85-95%) when compared to lexicon-based and machine learning approaches [29][30]. However, there remains a question left unanswered: can the same performance be achieved in the Sentiment Analysis for Chinese Text as well?

D. Limited research into Chinese Sentiment Analysis

Although there has been increasing amount of research into methods to recognize positive (favourable) and negative (unfavourable) sentiments toward specific subjects from online text, most of these efforts are directed towards Sentiment Analysis of English text, while the field of Chinese Sentiment Analysis remains nascent, if not underdeveloped. In fact, the latter deserves more attention, especially when one considers the sheer size of the Chinese e-commerce market, the expanding use of the Chinese language on the Web and availability of user-generated data across these platforms.

As of 2017, China is the world's largest and fastest growing e-commerce market, with 730 million internet users, nearly double the number of internet users in the whole of the United States. [31]. This paper hence affirms the importance of Chinese sentiment analysis a critical research field [32].

While there had been past papers written on Chinese sentiment analysis, the results yielded were relatively low and lacked the robustness and accuracy rates achieved by its counterparts for English Sentiment Analysis [25][32]. Furthermore, those methods tended to adopt the traditional machine learning approaches such as Support Vector Machines and Naïve Bayes Methods [33][34][35]. Some researches, in an attempt to utilize the well-developed resources of English Sentiment Analysis, even explored translating native Chinese text into English text and applying sentiment analysis to it [36]. However, these methods are tedious and do not achieve the optimal performance.

In recent years, there has been some progress made in research on deep learning techniques for usage in sentiment analysis of Chinese text. Today, Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) are the most popular forms of deep learning architectures used.

The former, RNN, was used to perform sentiment analysis for Chinese microblogging sites such as Weibo [37]. The latter, CNN, is another technique used on mining opinion summarization in Chinese microblogging sites and initially made popular due to its effectiveness in classifying images, was able to classify natural disaster events tweets with an accuracy of 89.47% [38].

In summary, these initial publications have shown that such deep learning techniques can be applied to Chinese text while maintaining its outperformance vis-à-vis traditional machine learning approaches and deserves more attention in expanding research into this field. Therefore, this paper seeks to create value-add to the nascent field of Chinese Sentiment Analysis; by presenting a new iteration of the RNN. Explicitly, this proposed iteration adopts the Gated Recurrent Unit (GRU) of the Recurrent Neural Network (RNN) architecture in Deep Learning for Chinese review sentiment classification. [39].

E. Sentiment Analysis using Recurrent Neural Network

First developed in the 1980s, Recurrent Neural Network (RNN) is a type of neural network best known for its ability to handle sequential patterns seen in text, whereby a word in a sentence holds a relationship with other words in the same sentence.

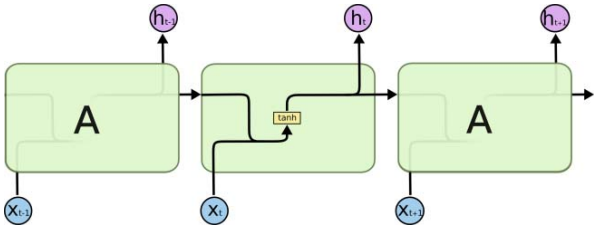

Fig. 2. Recurrent Neural Network with tanh activation function layer

Fig. 2 shows how an RNN works by applying the function (tanh) on a combination of its previous output (h_t-1), and a new input (x_t) used to predict output (h_t) at the time step, t. This process is done recursively for the length of the sentence. E.g. If the sentence consists of 5 words, hence, five sequences, the RNN will have 5 time-steps/layers.

The process of transferring the output of the first cycle together with the input of the next cycle enables the RNN to effectively "memorize" parts of the sentence and understand the relationship between words.

However, one weakness of RNN is that it faces the vanishing/exploding gradient problem, whereby gradients become either too small or too large during the training of the neural network [40]. This causes the network to have difficulty in learning/memorizing words that are further along in the sequence (sentence), i.e. long-term dependencies. Consequentially, the prediction accuracy gets weaker as the input sequence (sentence) increases in length, which is undesirable desirable in applications like sentiment analysis.

Researchers later designed two more advanced variations of the RNN, the Long Short-Term Memory and the Gated Recurrent Unit later designed by researchers, both with the same goal of mitigating the vanishing/exploding gradient problem allowing them to track long-term dependencies effectively.

F. Long Short Term Memory (LSTM)

The Long Short-Term Memory (LSTM) was first introduced in 1997 [41]. One of the key features of LSTM is that it consists of three gates, i.e. the input (i), forget (f), output (o) gates, and a memory cell (c) shown in Fig. 3.

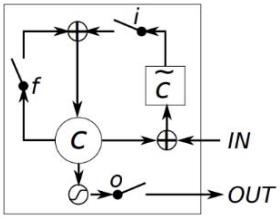

Fig. 3. Long Short-Term Memory Unit

With the introduction of the gates, this allows the LSTM to manipulate the contents stored within the memory cell. This significantly differs from the RNN unit, whereby at each time-step/layer, the material is completely overwritten. In the event LSTM senses an essential feature, it can pass this information over a longer distance, enabling it to capture long-term dependencies [42]. However, LSTM's main disadvantage lies with the long training time required, due to its temporal learning method which makes it difficult for parallel processing.

III. Proposed Approach for Sentiment Analysis

A. Sentence-level Sentiment Analysis

A review left by a consumer after an e-commerce purchase typically consists of 2 parts: The first part is quantitative, with a rating scale from 1 – 5. The second part is qualitative, in the form of a comment (a text body made up of one or more sentence) detailing the customer's experience and opinions about the product. When performing sentiment analysis on a review at the document level, i.e. the review in its entirety, the sentiment classification tends to correlate with the overall rating scale given.

TABLE I. Review Example

Review (Document Level)	Stars
首先，这款鼠标足够炫酷，侧键使用也十分方便，同时提供相应驱动来自行设计。无极滚轮和实时调	5

节dpi都是其亮点，在这个价位算是非常高性价比的了。但是，这款鼠标相比而言更加适合办公而非游戏，不加配重已经是我原来<brand x>鼠标的两三倍重量了，打了一下午的lol铂金局手腕就已经发酸，打吃鸡的时候因为重量的原因，感觉横移的灵敏度和精准度也不是非常的高。当然，作为一款办公鼠标，还是十分值得推荐的。好评。(Translation: Firstly, the mouse looks cool, its side button is also convenient to use, and the driver updates automatically. The scroll wheel and DPI adjustment are highlights, great features offered for this price point. However, this mouse seems more suitable for office than gaming, even without the additional weights, it is 2-3x the weight of my original gaming mouse. After playing an afternoon of games, my hand feels tired. While playing PUBG, because of the weight, I do not feel that the sensitivity and accuracy is very high. Definitely, as an office mouse, it is worth recommending.)	

Using the above example in Table I: The customer had given a 5-star rating for a product. By applying sentiment analysis at the document level, the review might have appeared to be a positive one on first glance. Such is often the situation in marketing departments, whereby marketers typically only monitor the average rating of individual products received. While one may conduct periodic scanning of 1-star rating reviews to gain a broad sense of product issues and dissatisfaction flagged out by customers, such a method fails to reflect the full picture, especially considering how only 7% of all reviews are given a 1-star rating [43].

TABLE II. SENTENCE SPLIT REVIEW WITH SENTIMENTS

Sentence Level	Sentiment
首先，这款鼠标足够炫酷，侧键使用也十分方便，同时提供相应驱动来自行设计。Firstly, the mouse looks cool, the side button is also convenient to use, and the corresponding driver updates automatically.	Positive
无极滚轮和实时调节dpi都是其亮点，在这个价位算是非常高性价比的了。The scroll wheel and DPI adjustment are highlights, features that are not usually offered at this price point.	Positive
但是，这款鼠标相比而言更加适合办公而非游戏，不加配重已经是我原来<brand x>鼠标的两三倍重量了，打了一下午的lol铂金局手腕就已经发酸，打吃鸡的时候因为重量的原因，感觉横移的灵敏度和精准度也不是非常的高。However, this mouse seems more suitable for office than game, without the additional weights, it is 2-3x the weight of my original gaming mouse. After playing an afternoon of games, my hand is tired. While playing PUBG, because of the weight, I do not feel that the sensitivity and accuracy is very high.	Negative
当然，作为一款办公鼠标，还是十分值得推荐的。好评。Definitely, as an office mouse, it is worth recommending.	Positive

The proposed method of splitting the review by sentence and applying sentiment analysis serves to provide business stakeholders and an additional parameter to filter data.

Table II illustrates how a review looks like after being broken down into four parts at the sentence level. By sieving out both positive and negative sentiments, it presents greater granularity on the review: While the mouse has exquisite design features and is priced reasonably (positive), the customer felt that it weighed heavier than other regular mice (negative) and would be more suited for work than gaming.

Armed with such breakdowns of consumers' feedback, this method allow businesses to be more in-tune to valuable emotions and feedback from customers about their brands and react accordingly: Marketing managers can make use of positive product features raved by customers and reposition marketing campaigns, while negative sentiments can provide product managers with ideas on future product refinements.

Given the greater depth of insights generated through sentence-based sentiment analysis as compared to document-based sentiment analysis, the former method was selected for this research.

B. Detailed Steps of Proposed Approach

Fig. 5 outlines the plan of approach for the preparation of the review corpus and training of the Gated Recurrent Unit for the purpose of Chinese Sentiment Analysis.

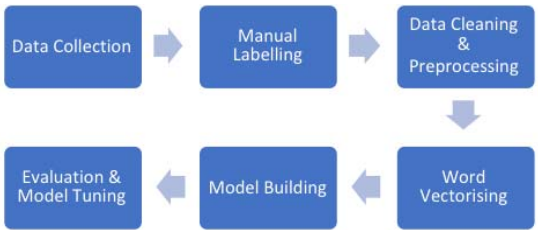

Fig. 4. Flowchart of Proposed Method

1) Data Collection

In order to prepare the dataset, 32,558 reviews (Document Level) were scraped from various Chinese E-commerce websites. These reviews were written by Chinese customers who had purchased the top 32 selling products of a consumer-electronics company over the last two years. Next, the reviews were split by its punctuation yielding 60,830 sentences.

Additionally, unwanted, duplicated and meaningless sentences were removed from the corpus as well. Notwithstanding, in order to reduce machine bias as much as possible, a well-distributed and extensive dataset was prepared which included expressions of happiness, disappointment, negation, adverbial and sarcasm, this ensured that the sentiment analysis model was comprehensively trained over a vast range of emotions and word usage, thus maintaining a relatively high degree of neutrality.

2) Manual Labelling

The 60,830 sentence level reviews were labelled manually by a native Chinese speaker who has over eight years of experience using social media and e-commerce websites and is familiar with the linguistic features of computer-mediated communication. The sentence-level sentiment analysis was formulated as a 2-class classification problem (positive/negative). Upon manual labelling, 60,830 sentence-level reviews resulted in 30,876 positive and 29,954 negative instances. This labelled corpus would serve as the dataset used to the train the model.

3) Data Cleaning and Preprocessing

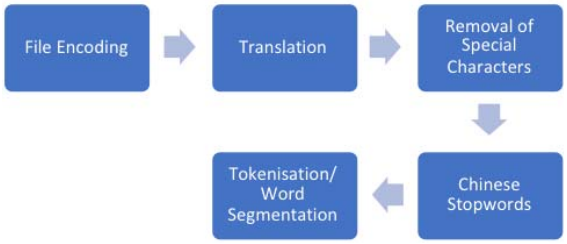

Fig. 5. Flowchart of Pre-processing Steps

Fig. 6 shows the data cleaning and pre-processing steps required to prepare the data before tokenisation. Detailed explanations of each stage are as follows:

a) File Encoding

UTF-8 is the most commonly used character encoding for the World Wide Web and accounts for 91.6% of all web pages. However, for the purpose of this project, GB 18030 (character encoding to read Chinese characters) was chosen as it delivers most robust performance regarding handling a wide range and variety of Chinese text (including simplified and traditional Chinese characters) over UTF-8.

b) Translation

There are currently two types of written Chinese characters: Traditional Chinese and Simplified Chinese. Traditional Chinese are original Chinese characters that do not contain newly created characters/character substitutions of the simplified Chinese writing system introduced in the 1950s. Simplified Chinese are standardised Chinese characters prescribed in the Table of General Standard Chinese Characters and contain fewer strokes per character.

In this project, all Chinese characters in the dataset were translated into simplified Chinese. By translating the dataset into a standardized format, i.e. Simplified Chinese, this would help the sentiment analysis model to learn the characters better and improve training accuracy. For instance, 数 (simplified) and 數 (traditional) may have the same meaning (count) but are written differently in simplified and traditional Chinese. However, the model does not understand the similarities and will process them as separate characters. This would adversely affect the model's accuracy should 数 be labelled as a positive sentiment, while 數 used in another sentence be tagged as a negative sentiment.

c) Removal of Special Characters

Special characters, e.g. *#@! does not add any meaningful information to the training of the sentiment analysis model. While an in-depth analysis revealed that special characters like emoticons were used in some reviews, the inclusion of special characters was deemed to convolute the integrity and results of the database potentially and was hence removed. This was also given that consideration that the project was on sentiment analysis, specifically on Chinese text. The removal also resulted in a reduction of the size of the training and test dataset.

d) Chinese Stopwords

The removal of stopwords and its impact on the prediction accuracy for sentiment analysis is a topic often debated by data scientists: Supporters argued that since stopwords do not contain any meaningful information, removal of such data could help to reduce memory overhead and noise, allowing the model to focus on more critical terms and hence improve prediction power. On the other hand, detractors believe that the neural network can learn the dependency information between words in a sentence. This will be an area for further validation for the project.

e) Word Segmentation

Word Segmentation, or otherwise known as tokenization of raw text is a standard pre-processing step for many NLP tasks. For English, Word Segmentation usually involves punctuation splitting and separation of some affixes like possessives. For Chinese, there are some characteristics of Chinese Text which differs from English Text, where it is scriptio continua, i.e. the style of writing is in a continuous script without spaces. Therefore, the step of splitting the Chinese text into a sequence of words is defined according to a word segmentation standard.

TABLE III. WORD SEGMENTATION (BEFORE AND AFTER)

Word Segmentation (Before)	Word Segmentation (After)
用习惯<brand x>的，总觉得微动太软了，直接拆了换了微动，其他挺不错，好看，手感也很好，适合手大的	用_习惯_<brand x>_的，总_觉得_微动_太软_了，直接_拆_了_换_了_微动，其他_挺不错，好看，手感_也 很_好，适合_手大_的

Jieba, a python package used for word segmentation, was chosen for this task. It works by first tokenizing the individual Chinese characters, then joining the tokens with spaces, before returning the complete sentence. Table III shows an review before and after word segmentation.

4) Word Embedding

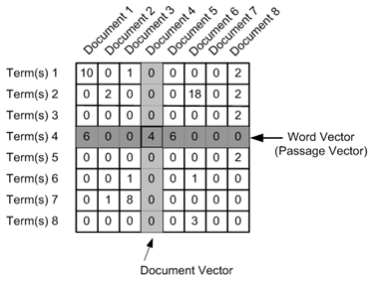

Fig. 6. Representative Image of a Count Vector

Most machine learning models, including neural networks, are unable to processing text in its raw form and require numbers as input. Therefore, word embedding is necessary to transform the dictionary of characters into a vector of continuous values. Count Vector is a frequency based embedding technique and is the most common technique used in language modelling for feature learning.

Fig. 7 is a representation of the output matrix of a count vector. The X-axis represents the terms, which are the dictionary of characters, each term is a unique character/s depending on the output from word segmentation (step v, pre-processing). Y-axis represents the document, which is the review sentence index. Each column is a review sentence. The count would represent the number of times; the term appears in the sentence.

5) Model Building

The proposed classification model is built using Gated Recurrent Unit (GRU). GRU is a simpler variant of LSTM but differs from the latter in 2 fundamental ways.

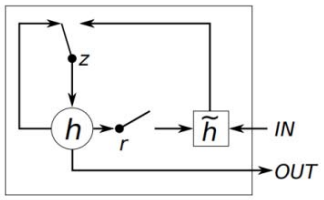

Fig. 7. Gated Recurrent Unit

Firstly, unlike the LSTM which has three gates, a GRU consists of 2 gates: an update (z) gate (which combines LSTM's forget and input gates) and the reset (r) gate, as shown in Fig. 7. Both gates regulate the flow of information within the unit: update gate (z) determines how much of necessary information are to be stored and passed on to the next output, while reset gate (r) determines and removes other irrelevant details. Secondly, while the LSTM unit makes use of a memory unit and output gate to control the flow of information to the next layer, the GRU bypasses the use of the memory cell and output gate. Instead, it exposes the full hidden content, h without control to the next unit/layer.

Despite its less complicated architecture, GRU outweighs LSTM in terms of benefits in sequence modelling: It is computationally more efficient, requiring shorter training time as compared to LSTMs. Next, due to lesser parameters, the code necessary for GRUs are easier to modify, maintain. Studies have even shown that GRU's performance that is comparable to LSTM not only in language modelling tasks but in polyphonic music modelling and speech signal modelling applications as well [44][42].

Given the aforementioned benefits, GRU was adopted as the chosen model used for Sentiment Analysis of Chinese Text in this paper.

6) Initial Baseline Model

The program was written in Python and implemented using Keras running on top of Tensorflow. The following parameters were used for the initial baseline model:

- Four-layer neural network
  - Embedding layer
    1. Input dimension: Total Vocabulary Size
    2. Output dimension: 256
    3. Input length: 20 Chinese characters
  - Two hidden layers (each with 256 neurons)
  - Output layer with 2 neurons with softmax function
- Activation function: None
- No Dropouts set for each hidden layer
- The Model was compiled with Adam optimiser, Categorical Cross-entropy loss functions, and accuracy metrics
- The Model was fitted with 32 batch size and 10 epochs
- Stopwords were excluded

The model is trained on 80% of the labelled data and validated (20%) for its accuracy. The model will subsequently be tested for its performance accuracy by using two different test datasets of product reviews. The baseline model achieved an accuracy of 80.15% on the validation dataset.

7) Evaluation

An optimal model is chosen using the validation accuracy obtained from the model training stage. Next, the selected model is used on two new datasets to verify its test accuracy. The first test dataset will be a balanced set (50:50) of positive and negative reviews. The second test dataset consists of an imbalanced number of positive and negative reviews (82:18). The latter set would be a more representative dataset of what the model would face more frequently after deploying the sentiment analysis model in production.

8) Optimisation

Machine learning algorithms usually come with a default set of parameters used for analysis. While it generally results in a well-performing model, it may lack the optimal configuration for the dataset and business problem. This is where optimisation (model tuning) comes into play, by obtaining the optimal values for the parameters or making changes in the pre-processing of the data of the chosen model, to enhance the model's capability to achieve the best validation accuracy possible.

The iterative process of tuning and evaluating the performance of a training model. It starts off by changing the hyperparameters of the model, or including/excluding specific pre-processing steps (usually through trial-and-error). Next, the model will be run on a training set. Following which, the performance will be evaluated on a validation set: This process is done iteratively until a parameter obtains the highest validation performance.

The GRU neural network classifier has several hyperparameters that can be tuned to improve its validation accuracy:

- Activation Function
- Dropout Rate
- Optimizer
- Input Length
- Number of Layers (Depth of Neural Network)

In the pre-processing stage, two training sets were prepared, one with Chinese stopwords be included and the other with Chinese stopwords excluded. Each set will be modelled using the same hyperparameters on GRU neural network to verify the hypothesis that removing stopwords would improve the accuracy of the training model.

9) Finalized Model

The final tuned GRU neural network classifier will have the following specified hyperparameters:

- Activation Function – tanh
- Dropout Rate – 0.9
- Optimizer – Adamax
- Input Length – 50
- Number of Hidden Layers – 2
- The Depth of Neural Network – 4

In the pre-processing steps of preparing the training set, Chinese stopwords will not be excluded as it contributes positively to the accuracy of the sentiment analysis model, as explained earlier. This final model achieved a validation accuracy of 84.43%.

10) Comparison to LSTM

In order to validate the potential improvement gained using GRU, a comparison with a state-of-the-art method in LSTM was chosen. A LSTM neural network classifier was ran with the similar optimal parameters of the GRU model against the training set. It achieved a validation accuracy of 83.5% (slightly lower accuracy, ~1%) and took 7% longer time in running each epoch (on CPU) compared to the GRU neural network classifier.

IV. Evaluation and Results

This section will elaborate on the metrics used to evaluate the performance of the sentiment analysis model. Next, the analysis on the results attained after the training model is used to predict on two types of test datasets (balanced and imbalanced). Finally, there will be a discussion on the approach taken by the sentiment analysis program in processing the document level reviews into sentence level reviews and assigning a sentiment (Positive or Negative).

A. Evaluation Metrics

The performance of the sentiment analysis model will be evaluated with the following metrics:

TABLE IV. Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$\text{Precision (P)} = TP / (TP + FP) \quad (2)$$

$$\text{Recall (R)} = TP / (TP + FN) \quad (3)$$

$$\text{F1 Score} = 2 * P * R / (P + R) \quad (4)$$

$$\text{Matthews Correlation Coefficient (MCC)} = (TP*TN - FP*FN) / [(TP+FP)(TP+FN)(TN+FP)(TN+FN)]^{0.5} \quad (5)$$

These are the most commonly utilized metrics used for the evaluation of classification modelling problems.

B. Sentiment Analysis Program Approach

The program will first split the document-level review by its punctuation into sentence-level punctuation and apply the pre-processing steps of data cleaning. Finally, the model will predict on the processed dataset by assigning each sentence level review a sentiment classification (Positive/Negative) and prediction confidence (0-100%).

C. Results on Balanced Data

The training model was predicted on a balanced test dataset of 50% positive and 50% negative reviews. Its performance was evaluated based on the following metrics. The solution developed achieved an accuracy score of 87.66%.

TABLE V. Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	644	157
Actual Negative	49	819

TABLE VI. Confusion Matrix

Accuracy	Precision	Recall	F1 Score	MCC
87.66%	92.93%	80.4%	0.8621	0.7579

The solution developed achieved an accuracy score of 87.66%.

D. Results on Imbalanced Data

These are the most commonly utilized metrics used for the evaluation of classification modelling problems. The training model should not only be tested on a balanced dataset but also on an imbalanced dataset to ensure model robustness. This is because an imbalanced dataset is more reflective of real-world scenarios when reviews tend to be more positive than negative [43]. Thus, the training model was predicted on an imbalanced test dataset of 82% positive and 18% negative reviews. Its performance was evaluated based on the following metrics:

TABLE VII. Confusion Matrix

	Predicted Positive	Predicted Negative
Actual Positive	1789	276
Actual Negative	46	408

TABLE VIII. Confusion Matrix

Accuracy	Precision	Recall	F1 Score	MCC
87.2%	97.49%	86.63%	0.9174	0.6611

The solution developed achieved an accuracy score of 87.9%. This model was subsequently deployed into production for users to perform sentiment analysis predictions on Chinese reviews datasets.

V. Discussion

This section will discuss the value achieved by this paper. Firstly, this paper has contributed to the research field of Chinese Sentiment Analysis by creating an extensive training corpus of 60,830 and two test corpus (balanced and imbalanced) of 2171, 2519 labelled Chinese reviews in the domain of consumer electronics.

Secondly, by adopting highly efficient deep learning technique of Gated Recurrent Unit (GRU) neural network, this paper had achieved an accuracy rate of (87.66% - balanced, 87.9% imbalanced). This result is mostly in line with other state-of-the-art deep learning techniques such as Long Short-Term Memory (LSTM) and Convoluted Neural Networks (CNN).

Thirdly, a new sentence-level analysis was proposed in this paper, which can potentially provide more granular and in-depth insights on consumer sentiments to business stakeholders.

Fourthly, this paper has demonstrated the positive benefits of using deep learning:

- Deep learning allows for gradual improvements to the training model through additions of unobserved types of reviews or emoticons into the corpus and subsequent re-training;
- Deep learning allows for the sentiment analysis model to be applied to other languages, with the condition that the respective language corpus is obtained for training the mode;

- Deep learning has the ability of learning negation words that could reverse the polarity of sentiments well, simply by including negation examples with the appropriated labels in the dataset and training the model;

Lastly, while the paper had illustrated the usage of character tokenisation for sentiment analysis, it further highlights the potential of character tokenisation for areas of extensions, including Word Cloud, Term Frequency-Inverse Document Frequency (TF-IDF), Aspect Extraction and Topic Modelling.

VI. Conclusions and Future Work

This paper contributed in supplementing the void of research into Chinese sentiment analysis by improving upon the existing sentiment analysis techniques (lexicon/machine learning based techniques) in the proposal of the state-of-the-art deep learning technique of using Gated Recurrent Unit (GRU) neural network in the classification (positive/negative) of consumer electronics reviews.

Another area worth noting is the proposed method of giving increased granularity to the review by splitting the review by sentence which helps provide stakeholders an additional dimension of insights to the existing data.

The paper also outlined a standardised approach of pre-processing Chinese Text as input for any Deep Learning Model and created a brand-new corpus of labelled reviews data within the domain of consumer electronics.

Finally, this paper has accomplished the technical objectives and business considerations set at the onset of this project, achieving experimental results of 87.66% on a balance test dataset and 87.9% on an imbalanced test dataset. In summary, the results attained ascertain that the proposed approach is efficient, scalable, robust and highly accurate.

As much as the paper tries to be extensive in its coverage of the nascent area of Chinese Sentiment analysis using Deep Learning Techniques, there remain other potential areas for further work that merit consideration. An interesting area for further work can be found in the pre-processing step of word embedding, there appears to be more advanced methods of converting text into a vector space such as Word2Vec, GloVe (Global Vectors for Word Representation) and fastText. These methods could potentially be an improvement to the current word encoding method used for this paper. This is because such methods can capture the semantic relationships between each character and its relation to the entire character corpus.

Acknowledgment

This project was done in a collaboration with Logitech Europe S.A. and National University of Singapore, Business Analytics Centre. Jun Sheng, Lee would personally like to thank Mr. Denis Zuba and Associate Professor Pang Yan for their valuable inputs and guidance during the course of this research paper.

References

[1] K. Burke, "73 Texting Statistics That Answer All Your Questions," 24-May-2016. [Online]. Available: https://www.textrequest.com/blog/texting-statistics-answer-questions/. [Accessed: 20-Jul-2018].

[2] Statista, "Facebook users worldwide 2018," *Statista*, 2018. [Online]. Available: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/. [Accessed: 21-Jul-2018].

[3] A. Singhal, "Google Search Statistics - Internet Live Stats," 2012. [Online]. Available: http://www.internetlivestats.com/google-search-statistics/. [Accessed: 21-Jul-2018].

[4] IDC, "Executive Summary: Data Growth, Business Opportunities, and the IT Imperatives | The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things," Apr-2014. [Online]. Available: https://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm. [Accessed: 13-Jun-2018].

[5] C. Schneider, "The biggest data challenges that you might not even know you have," *Watson*, 25-May-2016. [Online]. Available: https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/. [Accessed: 13-Jun-2018].

[6] K. Taylor, "Want to Reach Millennials? This Is How They Spend Their Time. (Infographic)," *Entrepreneur*, 09-Oct-2014. [Online]. Available: https://www.entrepreneur.com/article/238294. [Accessed: 21-Jul-2018].

[7] L. Wright, "The 50 User Generated Content Stats You Need to Know," *Medium*, 27-Apr-2017. .

[8] Newswire, "Newswire | Consumer Trust in Online, Social and Mobile Advertising Grows | Nielsen," 10-Apr-2012. [Online]. Available: http://www.nielsen.com/us/en/insights/news/2012/consumer-trust-in-online-social-and-mobile-advertising-grows. [Accessed: 21-Jul-2018].

[9] I. Varlamis, M. Eirinaki, and D. Proios, "TipMe: Personalized advertising and aspect-based opinion mining for users and businesses," in *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2015, pp. 1489–1494.

[10] P. D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," *arXiv:cs/0212032*, Dec. 2002.

[11] P. J. Stone, D. Dunphy, M. S. Smith, and D. M. Ogilvie, "The General Inquirer: A Computer Approach to Content Analysis," *The MIT Press*, 1966. [Online]. Available: https://mitpress.mit.edu/books/general-inquirer. [Accessed: 21-Jul-2018].

[12] R. Tong, "An operational system for detecting and tracking opinions in on-line discussions," *Work. Notes SIGIR Workshop Oper. Text Classif.*, pp. 1–6, 2001.

[13] V. Hatzivassiloglou and K. R. McKeown, "Predicting the Semantic Orientation of Adjectives," p. 8, 1997.

[14] J. M. Wiebe, "Learning Subjective Adjectives from Corpora," p. 6, 2000.

[15] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," p. 10, 2004.

[16] M. Taboada, C. Anthony, and K. Voll, "Methods for Creating Semantic Orientation Dictionaries∗," p. 6, 2006.

[17] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, 2002, pp. 79–86.

[18] J. Bartlett and R. Albright, "152-2008: Coming to a Theater Near You! Sentiment Classification Techniques Using SAS® Text Miner," p. 9, 2008.

[19] E. Boiy, P. Hens, K. Deschacht, and M.-F. Moens, "Automatic Sentiment Analysis in On-line Text," p. 12, 2007.

[20] A. Kennedy and D. Inkpen, "Sentiment Classification of Movie and Product Reviews Using Contextual Valence Shifters," p. 14, 2006.

[21] P. Chaovalit and L. Zhou, "Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches," in *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, 2005, pp. 112c-112c.

[22] B. Pang and L. Lee, "Opinion mining and sentiment analysis," p. 94, 2008.

[23] A. Aue and M. Gamon, "Customizing Sentiment Classifiers to New Domains: a Case Study," p. 7, 2005.

[24] R. Moraes, J. F. Valiati, and W. P. Gavião Neto, "Document-level sentiment classification: An empirical comparison between SVM and ANN," *Expert Syst. Appl.*, vol. 40, no. 2, pp. 621–633, Feb. 2013.

[25] S. Liao, J. Wang, R. Yu, K. Sato, and C. Zixue, "CNN for situations understanding based on sentiment analysis of twitter data," 2016. [Online]. Available:

https://reader.elsevier.com/reader/sd/426B35B76E6941236F8CF722A3C1D564597824906926285C920CD562B0EC8897918F41AA2692277830E73BE1CF88DB70. [Accessed: 21-Jul-2018].

[26] X. Sun, C. Li, and F. Ren, "Sentiment analysis for Chinese microblog based on deep neural networks with convolutional extension features," *Neurocomputing*, vol. 210, pp. 227–236, Oct. 2016.

[27] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," 2009, pp. 1–8.

[28] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *ArXiv13013781 Cs*, Jan. 2013.

[29] R. Socher *et al.*, "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA, 2013, pp. 1631–1642.

[30] C. dos Santos and M. Gatti, "Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts," in *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin, Ireland, 2014, pp. 69–78.

[31] R. Flannery, "What Makes China's Internet Growth So Fast And Volatile?," 2017. [Online]. Available: https://www.forbes.com/sites/russellflannery/2017/10/10/what-makes-chinas-internet-growth-so-fast-and-volatile/#7dface394852. [Accessed: 22-Jul-2018].

[32] H. Peng, E. Cambria, and A. Hussain, "A Review of Sentiment Analysis Research in Chinese Language," *Cogn. Comput.*, vol. 9, no. 4, pp. 423–435, Aug. 2017.

[33] Y. H. H. Alani and D. Zhou, "Exploring English Lexicon Knowledge for Chinese Sentiment Analysis," p. 8, 2010.

[34] S. Tan and J. Zhang, "An empirical study of sentiment analysis for chinese documents," *Expert Syst. Appl.*, vol. 34, no. 4, pp. 2622–2629, May 2008.

[35] D. Zhang, H. Xu, Z. Su, and Y. Xu, "Chinese comments sentiment classification based on word2vec and SVMperf," *Expert Syst. Appl.*, vol. 42, no. 4, pp. 1857–1863, Mar. 2015.

[36] X. Wan, "Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis," 2008, p. 553.

[37] J. Liang, Y. Chai, H. Yuan, H. Zan, and M. Liu, "Deep learning for Chinese micro-blog sentiment analysis," *J. Chin. Inf. Process.*, vol. 28, no. 5, pp. 155–161, 2014.

[38] Q. Li, Z. Jin, C. Wang, and D. D. Zeng, "Mining opinion summarizations using convolutional neural networks in Chinese microblogging systems," *Knowl.-Based Syst.*, vol. 107, pp. 289–300, Sep. 2016.

[39] K. Cho *et al.*, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *ArXiv14061078 Cs Stat*, Jun. 2014.

[40] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

[41] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[42] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," *ArXiv14123555 Cs*, Dec. 2014.

[43] Datafiniti, "The Connection Between Grammar and Product Reviews," *Datafiniti*, 06-Feb-2018. [Online]. Available: https://datafiniti.co/grammar-of-online-reviews/. [Accessed: 25-Jul-2018].

[44] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," *ArXiv14090473 Cs Stat*, Sep. 2014.