Twitter- a new pathway to access product innovation ideas - Can machine learning help PepsiCo identify innovative ideas in User-Generated Content platforms?

Author: Yucheng Chen
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands

ABSTRACT,
In the last decades, fulfilling ever-changing customer needs in this changing market environment has been chased by multiple commercial entities, which has created enormous opportunities for machine learning-based User-Generated Content (UGC) analysis. For information extraction tasks from unstructured UGC, natural language processing (NLP) approaches based on machine learning are becoming increasingly popular for information extraction tasks from unstructured UGC. The objective of this paper is to quantify and enhance Text CNN's text classification performance on tweet datasets generated by the Twitter keywords filter. It will discuss the primary performance of Text CNN and the problem this model faces during dealing with the task of identifying numerous tweets. Through discussing the different performance metrics, it suggests using a macro F1 score as a baseline and concentrating on improving the recall of the class "without CNs". Then by identifying problems like imbalanced datasets, this paper discusses previous solutions.
Furthermore, this research also intends to analyse the model's application to PepsiCo's flavour innovation processes through the extracted information. It also suggests that PepsiCo may empower the cost-efficiency of new flavours market research via adopting the machine learning-based UGC analysis. Lastly, recommendations based on observations will be provided for further research on the application of machine learning models in the beverage industry.

Graduation Committee members:
First supervisor: Dr Dorian Proksch
Second supervisor: Dr Tim G. Schweisfurth

Keywords
Data mining, user-generated-contents analysis, customer need elicitation, data-driven innovation strategy, beverage industry, Text CNN.

1. INTRODUCTION

Customer needs (CNs) are often defined with words or phrases that describe what consumers seek to acquire from items. How to fulfil ever-changing customer needs in this changing market environment has been discussed for an extended period, as customer need is considered a vital asset for business innovation. Corporates expect to satisfy the market gap to obtain competitive advantages and acquire a dominant market position. Traditionally, firms depend on interviews and focus groups to research customer needs, which can be time-consuming and costly and constantly lead to a late response to business opportunities. In contrast, User-generated Content (UGC) furnishes instant indication of customer preferences, enabling a cost-efficient way to mine market insights (Timoshenko & Hauser, 2019). With the rise of social media, for instance, Twitter, Facebook, Instagram and YouTube, the values of UGC are increasingly emphasised by entrepreneurs for seeking benefits (Fischer& Reuber, 2010). To effectively derive information from big data in online communities, which is a critical source of creative ideas for product innovation, text mining and machine learning (ML) are widely used to find hidden thoughts behind substantial quantities of text (Christensen, Nørskov, Frederiksen & Scholderer, 2016).

However, although numerous industries increasingly concentrate on innovation, the food industry has a low level of innovation since it is mature and inching (Bigliardi, Ferraro, Filippelli & Galati, 2020). Nevertheless, in recent years, various social, economic, and technical developments in the food industry have enforced significant changes in food demand and supply chain organisation. It makes innovation an unavoidable activity critical to a company's success. A successful business innovation example is the accelerated development of Genki Forest in the Chinese beverages market. Genki Forest has seized the market needs for sugar-free, fat-free and calorie-free beverages and developed a series of new flavours, for instance, cucumber, sakura and so on. The launching of unseen beverages enables them to achieve a 334% compound growth rate (Xinhua, 2021). The development of new brands reduces the market share of oligarchs in the beverage industry. This essay intends to investigate the leading beverage company, in this case, PepsiCo, using social media analysis through ML algorithms to drive the product innovation process for coping with the attack from new rivals.

1.1 Research objective

The current innovation strategies from beverage companies emphasise the testing perspective, which indicates that the cautious innovations due to the cannibalisation effect (The reduction in companies' sales, profits or market share caused by the launch of a new product or service from the same company [Lomax & McWilliam, 2001]) enforces corporates to launch extended products in a small-scale. However, traditional market research for innovative products requires a significant amount of time and money. Thus, this research aims to explore a new pathway for providing product innovation ideas for beverage companies based on User-Generated Content from social media platforms.

1.2 Research problem

- *"Can machine learning help PepsiCo identify innovative ideas in User-Generated Content platforms?"*

This research problem investigates whether PepsiCo can use a machine-learning algorithm to obtain customers' voices on social media platforms (Twitter) to gain competitive advantages or survive in the beverage market with fierce competition. Since PepsiCo has an international business, the rising local brands use creative product innovations to erode Pepsi's market share. Thus, to satisfy the changing global customer needs, new methods of automatic collection of customer needs are needed when the traditional market research methods may lead to missed business opportunities.

1.3 Sub question

- *"To what extent can social media analysis based on machine learning algorithms affect the innovation process?"*
- *"Can the machine learning algorithm be used to identify customer needs in social media for inventing new flavours of beverages?"*
- *"How could PepsiCo use machine learning to identify new flavours that may fulfil potential market gap?"*

2. LITERATURE REVIEW

The development of data-driven innovation (DDI) has influenced the current business strategies and paradigms (Lies, 2019) and "led to the emergence and development of new products, business models, and opportunities in a digital ecosystem" (Saura, Ribeiro-Soriano & Palacios-Marqués, 2021). The digital ecosystem consists of digital markets where the information generated from user actions is stored in the form of data (de Camargo Fiorini, Roman Pais Seles, Chiappetta Jabbour, Barberio Mariano & de Sousa Jabbour, 2018). This section intends to clarify key concepts related to UGC analysis and the machine learning approach in data-driven innovation processes. Firstly, it will illustrate the significance of fulfilling customer needs for corporates and then identify the traditional research pathway of conducting customer needs. Furthermore, the relevance between UGC analysis and customer insights will be discussed. Lastly, related works on classifying relevant data and identifying the importance of extracted information will be addressed in this paper.

2.1 Customer needs and User-Generated contents

User-Generated Content is defined by Krumm, Davies & Narayanaswami (2008) as data, information or media that contribute by people on online platforms. UGC contains much richer customer opinions and feedback, providing a new method to observe customer needs (the problems buyers want to remedy by purchasing a product or service [Rahman & Safeena, 2016]); thus, it is increasingly attractive for corporates.

2.2 Customer needs, innovation and company performance

Based on Majava et al. (2014), customer needs are generated from the problems customers have and correlate with customers' values and behaviour; how to fulfil customer needs is vital but also challenging for firms. Chong and Chen (2009) also consider that the success rate of new product development (NPD) is closely correlated with fulfilling customer needs. They propose that market uncertainty adversely affects projects' success rate; however, NPD projects may obtain a higher success rate by defining unfulfilled customer needs precisely. Shane (2009) also recognises customer needs as a critical step in the innovation processes; the Voice of Customers, as an initial step for multiple innovation models, requires the input of CNs, which leads to a prioritised needs hierarchy that helps cross-function teams to make trade-offs decisions. Understanding customer needs also consistent with the essence of customer relationship management (CRM) based on Stringfellow, Nie and Bowen's (2004) research. It states that the deployment of CRM through leveraging the knowledge of CNs can have a significant effect on firms' bottom-line performance.

The previous statement suggests that understanding CNs has positive impacts on performing innovation and enhancing firms' performance, which motivates corporates and researchers to investigate the composition of CNs. Nevertheless, gaining customer insights is ambitious; thus, various theories were developed by academics to explore customer needs' nature. Shillito (2000) illustrated that there are three tiers of CNs, which are features, consequences, and desired end-states. Shane (2009) further discussed three aspects of CNs. Features are frequently used to describe a product or service. For instance, a cell phone has a screen, battery, embedded memory, et al.; it concentrates on the nature of the short-term change and is commonly adopted by the incremental changes. Consequences arise from owning or using the goods or service (The cell phone, for example, is easy to operate and has a long battery life); it is primarily used to explain what the consumer desires and is often more emotive in nature. Customers' fundamental intents and goals are the desired end-states; as a result, they are more long-term and abstract (for example, a cell phone helps me record wonderful experiences).

2.3 Traditional Methods to Identify Customer needs

With understanding the essence and significance of CNs, various channels of collecting methods are proposed by previous works. Majava et al. (2014) stated that (a) interviews, (b) observations, (c) focus groups, (d) becoming a user, (e) customer advisory boards, (f) websites, (g) panels, (h) groups brainstorming, (i) innovation summits, (j) customer integration into a product development team, (k) customer dialogues, (l) ethnography, (m) identifying lead users, and (n) market surveys are approaches frequently employed for acquiring CNs. Salminen, Jung & Jansen (2021) further explain the four methods involved. Interviews adopt open-ended questions to capture responses that reflect a user's feelings about a product or service. The sample size for each study varies based on the project's goals and resources. According to Griffin and Hauser (1993), one-hour interviews with 20- 30 participants can elicit 90-95 per cent of user requirements. While some studies may have more than 30 participants, extended interviewing has negative results; active interaction with users and an independent evaluation of user activities are examples of observation. In this approach, people's actions are more crucial than what they say about their needs. As a result, it does not require users to be aware of their desires. Ethnography is a type of observation that entails extensive immersion in a natural context (such as supermarkets, hospitals, workplaces, and schools) better to understand user habits and behaviours in real-world circumstances.

Salminen, Jung & Jansen (2021) define a focus group as a pathway to summarise diverse perspectives of user values at one point of time in customers' discussion. The survey refers to various probability sampling methods; users are asked to answer open-ended or closed-ended questions by researchers. Although the manual methodologies have evolved to be mature, several challenges constrain the research on CNs. Salminen, Jung & Jansen (2021) mentioned that, due to the nature of traditional methods, the scale and sample size of the above techniques are significantly limited. Although techniques like in-depth interviewing have a higher possibility of capturing latent needs, mass attempts by automagical collection may be more likely to catch. Schaffhausen & Kowalewski (2015) and Kühl, Mühlthaler & Goutier (2019) also noted budgetary constraints and time boundaries of manual collection. The manual qualitative research requires subject-matter experts to involve in all stages, while each step may require weeks-long field studies, which leads to a more extensive duration of research. Moreover, Wang et al. (2018) indicated the human bias in traditional methods. For instance, difficulties in expressing needs refer to the situation where users cannot articulate their intentions and preferences to others, leading to vague or misleading language in conveying their requirements. Due to these constraints, researchers are dedicated to developing automatic methods for user needs detection to fulfil the demands of large-scale market research.

2.4 UGC Analysis in identifying customers' needs

The traditional collection methods of CNs have contributed to customer research for an extended period. However, the customer needs have become more dynamic-orientation due to the reduced product lifecycle and globalised business environment (Jeong, Yoon & Lee, 2019). Chong and Chen (2009) created a scenario where the customer needs shift significantly between the design specifications frozen and the market introduction period; the business outcome of this scenario is suggested to be negative. Apart from the dynamic customer needs, the trend of social media platforms where users express their thoughts, feelings, and complaints about items and services (Liu, Jiang & Zhao, 2019) also accelerates the development of user-generated content analysis.

As a rich channel of data sources, UGC consists of enormous data. According to the critical-mass theory discussed by Prasarnphanich and Wagner (2009), if there is a sufficiently high number of supporters of a concept, technology, innovation, or social system, its adoption of them will be self-sustaining. In this case, the UGC provides constant support for ideas on social media platforms, which creates an ecosystem beneficial for both customers and corporates. However, as discussed in the previous part, the traditional pathways are time-consuming and poor performing in big data. The issue of information overload (when the cognitive processing capacity of a digital system is exceeded by the data needed to be processed) may occur and limit the quality of the outcome (Kaufhold, Rupp, Reuter & Habbdank, 2020).

2.5 The commercial values of fulfilling customer need in the beverage industry

The key to beverages that make customers like and prefer are the sensory property, for instance, taste, aroma, flavour, texture, appearance and sound (Tuorila, 2007). Highly distinct flavours also give a demonstrable basis for higher price points, increasing customer value and outperforming the competition ("Flavor as a Business Building Strategy", 2006). Although Brand recognition and product testing are essential, product satisfaction leads to repurchase. Consumers will not repurchase a food product if they dislike it, regardless of branding; hence, depending solely on the brand image to sell beverages is risky (Kemp, 2013).

As a result, understanding customers and including them in the innovation process to build products they want to consume, utilise, and enjoy is critical to increasing revenues and lowering risks for beverage corporates (Kemp, 2013).

2.6 Previous works on the machine learning approach in UGC analysis

Scholars have developed various models to filter out customer values to handle enormous data. Salminen, Jung & Jansen (2021) have roughly concluded these models into two types, (a) supervised ML, which involves manual labelling of training data, and (b) unsupervised ML, where no data is labelled, algorithms are used to detect and show data patterns.

This paper would consider text classification mainly include two steps, topic modelling and classifier choosing. In dual-topics classification, Coussement and Van den Poel (2008) used term-vector weighting to formulate a vector matrix that includes the weight of each token in each topic. Then, they reduced the

dimensions through Singular Value Decomposition. The process of classification was conducted by the Adaboost algorithm to distinguish the linguistic style of emails. The topic modelling in dual-topics classification only contains two topics that may not be well-performing in texts with multiple topics. Thus, researchers started to adopt different algorithms.

For papers with multiple topics to classify, the manually coded lexicon is commonly adopted for sentiment analysis which determines the specific domains of text and then assigns text into set topics through various classifier algorithms, for instance, KNN (Liu, Jiang & Zhao, 2019), K-means (Lee & Bradlow, 2011), Bayes classifiers, Support Vector Machines, random forests and CNN (Timoshenko & Hauser, 2019). Even though these models have enabled a quick and precise selection of information, UGC contains uncertain topics that cannot be determined before analysing, which may limit the performance of the above models. Thus, in Jeon., Yoon and Lee's (2019) research, Latent Dirichlet Allocation is used to identifying potential topics automatically.

3. METHODOLOGY

Since the UGC contains massive unstructured data, thus, it needs to be processed via multiple technics to present helpful information. This section intends to illustrate the methodology adopted by this research. As figure 1 shows, the data channel will be explained first with the data sources and search restrictions. Then, the data filter out noises documents that do not relate to the topics. Data labelling is the critical step that provides train sets and cross-validation sets that help improve the quality of machine learning models. Lastly, several natural language processing approaches will be discussed to extract the outcomes which are customer needs in this research.

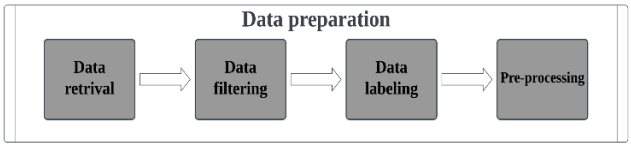

Figure 1. Data preparation steps.

The UGC, as defined previously, is the data that people on online platforms contribute. As one of the most popular social media platforms, Twitter is an attractive data source. Its information is accessible and can be analysed using data mining techniques to extract and process information provided by users worldwide (Massoudian, 2016).

This research expects to use keywords like "soda", "soft drink", "lemonade", and "cola" with mentioning "flavor" or "flavour" as selecting criteria to increase the correlations between selected tweets and topics. The retrieval process would provide 10,000 tweets posted within a year (30.03.2021-01.04.2022) that correlate with customer needs for new flavours of beverages. This research will utilise the free access API for researchers from Twitter to crawl the tweets that meet the requirements. API stands for Application Programming Interface. In this case, access to the API of Twitter means that we can access selected tweets and download them into various formats, which in this paper will be stored in an xlsx file.

3.1 Data filtering

The UGC is typically described as unstructured data, as mentioned above; moreover, on Twitter, global users can post any contents that follow the community regulation, which means that these tweets may include posts in non-English languages, which is useless for the research. Thus, non-English posts included in the dataset will be eliminated. Except for the language limitation, the commercial posts may also negatively impact model training. The commercial posts refer to the tweets from either official accounts or retailers' accounts for only marketing purposes that cannot represent latent needs, which commonly carry keywords like "sale", "win", "sales", "boost", and "learn more". Also, posts with URLs are excluded since most posts with URLs cannot effectively reflect CNs (Kuehl, Scheurenbrand & Satzger, 2020).

Moreover, the retweet function may cause the number of duplicated tweets to increase; thus, only regular tweets will be considered, and the repeated tweets will be deleted while retaining the oldest version. Besides, short Content may also be regarded as noise in this step since it may not be able to carry sufficient information related to this topic. Figure 2 below will present the number of tweets remaining in the dataset.

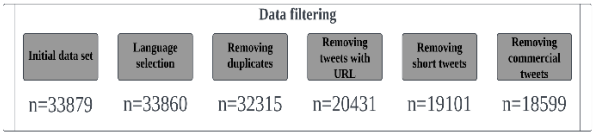

Figure 2. Amounts of the tweet in steps of data filtering.

3.2 Data labelling

After filtering out useless data, the remains needed to be labelled into two categories: with or without customer needs, for building supervised machine learning models. The tweets with CNs should express a strong willingness toward specific flavours. Following this standard, the below tweets will be regarded as unrelated, which will be indicated as "user need = 0" in the dataset. The labelling will be completed manually, which means that around 60% of the remaining data will be used as training data, and 40% of data will be used as test sets and validation sets, which may prevent overfitting (accurate in training data but not in testing data) or underfitting model (the accuracy of this model is low) (VanderPlas,2016). However, all tweets will be manually labelled to validate the model's performance.

3.3 Pre-processing

Transforming tweets into a standard format for subsequent processing with a filtered dataset is essential to training a precise model. The natural language processing steps are typically adopted to process information with human languages; it includes sentence segments, word tokenisation, stemming, lemmatisation, and stop word removal ("NLP Tutorial - Javatpoint", n.d.). Transforming tweets into a standard format for subsequent processing with a filtered dataset is essential to training a precise model. In this research, since the text needed to be processed is tweets from multiple users, which may include user names, for instance, "@seventhexit Depends if you just mean the soda flavor or the freeze!". The user name included in the text cannot provide sufficient needs but may indicate the categories of the product this tweet relates to; thus, user names will be deleted in the data pre-processing step instead of data filtering.

The next step is the sentence segments which break a whole paragraph into several sentences, while the word tokenisation breaks sentences into different words. However, word limitations exist for tweets; thus, the sentence segment will not be performed. Then, stemming normalise words into their root form; for instance, "loves" will be recognised as "love"; lemmatisation is highly close to stemming; it "groups different inflected forms of the word", which, for example, changes the word "better" into "good" ("NLP Tutorial - Javatpoint", n.d.). The last step, stop words removal, will delete the stop words with

the tokens set; the words like “he”, “she”, and “be” normally do not carry any information but only serve the linguistic purpose. Moreover, the occurrence of stop words is relatively high, which may disturb the analysis. The above steps will be conducted via Jupyter Notebooks for Python with modules like “pandas” (McKinney et al., 2010), “NLTK” (Bird, Klein & Loper, 2009) and so on. This research intends to train a machine learning model to identify potential customer needs on Twitter and extract the possible product innovations ideas. By manually specifying the pre-processed tweets data into “contains customer needs” and “does not contain customer needs” by using around 30% of data, a Text CNN model is expected to be trained to identify the tweets carried customer needs.

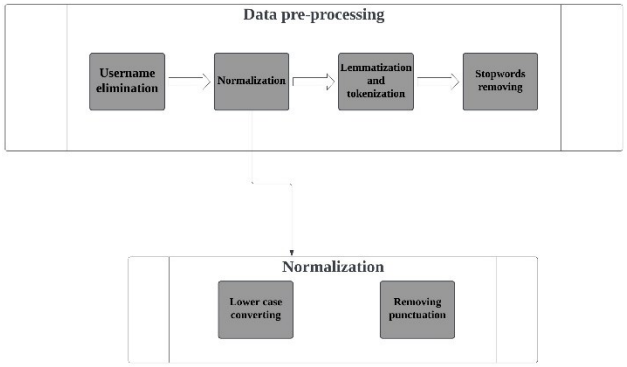

Figure 3. Method model of data pre-processing.

3.4 Manual extraction and categorisation

After labelling tweets containing customer needs, the machine learning classification method discussed above could not extract the customer needs from the tweets. Moreover, the machine learning pathway cannot identify tweets with customer needs, in this case, the beverage flavour ideas. Meanwhile, the language usage in tweets is various. According to Boot, Tjong Kim Sang, Dijkstra and Zwaan (2019), The length of utterances in spontaneous verbal communication is normally unrestricted. However, the length of tweets is restricted to 240 characters; this character-reduction method is called textisms. Textisms minimise the number of characters used without misleading the message’s inclinations. Examples of textisms are acronyms, emoticons, accent stylisations, nonconventional spellings, homophones, shortenings, contractions, and punctuation absence. While in this case, for instance, users may adopt “blueb” to refer to blueberry, which may be misleading for the machine learning algorithm. Thus, the extraction and categorisation of product innovation ideas should be completed via human force.

3.5 The convolutional neural network for text (Text CNN) for classification

In this research, the machine learning method is expected to realise the goal of text classifying numerous tweets into two categories: “with customer needs” and “without customer needs”. Over recent years, Convolutional neural networks (CNNs) have attracted much attention in the field of text categorisation in recent years due to their impressive performance in a variety of contexts (Guo, Zhang, Liu & Ma, 2019). While in the field of text classification, Text CNN is one of the most commonly used CNN models. Based on Kim (2014), The architecture of Text CNN consists of four parts, input layers (word embedding), convolutional layers, max-over-time pooling and a fully connected layer.

Word embedding is referred to as “mapping each word into a word vector” (Guo, Zhang, Liu & Ma, 2019), which means that each word is represented as real-valued vectors in a predefined vector space. While Word2Vec (Mikolov, Sutskever, Chen, Corrado & Dean, 2013) and Global Vectors (GloVe) (Pennington, Socher & Manning, 2014) are the most commonly used word embedding algorithms for converting words into low-dimensional dense vectors during the last few years.

According to O’Shea and Nash (2015), the convolutional layer aims to determine the output of neurons connected to local regions of the input by calculating the scalar product between their weights and the region connected to the input volume. Then, rectified linear unit (also known as ReLu) tries to apply an ‘elementwise’ activation function like “sigmoid” to the output of the preceding layer’s activation (O’Shea & Nash, 2015).

The pooling layer will then execute down sampling along the input’s spatial dimensions, lowering the number of parameters in activation, and the fully-connected layers will try to generate class scores from the activations used to classify the data (O’Shea & Nash, 2015).

4. RESULT

4.1 The setting of the Text CNN model

Description	Values
input word vectors	GloVe
filter region size	(3,4,5)
numebr of output filters	512
activation function	ReLU
pooling	1-max pooling
dropout rate	0.2
activation function	softmax

Table 1. Model configuration.

The model configuration of Text CNN applied to the tweets dataset is shown in table 1. In this case, GloVe transforms the input layers into word vectors. Then, three convolution layers are followed, with the filter windows of 3, 4, and 5. Each layer carries 512 output filters; the outcomes of each convolution layer will be activated by the function ReLu as the inputs for the max-pooling layer. After the flatten function, the outputs from concatenated layer (the layer which joined the outcomes from three max-pooling layers) are converted into dual dimensions. Subsequently, as figure 4 shows, there are two fully connected layers to generate the test score to improve the model performance. Moreover, to prevent overfitting, the dropout rate is set as 0.2. After constructing the Text CNN model, the network has 2.82 million trainable params, with the AdaDelta optimiser, a batch size of 64, epochs of 10 and a learning rate of 1e-5.

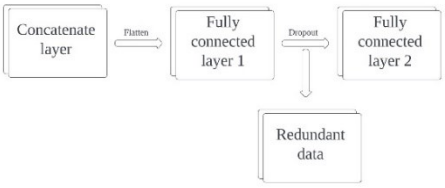

Figure 4. Structure of fully connected layers.

Before training the model with pre-divided datasets, it is crucial to notice that the number of labels is imbalanced; the tweets with CNs only represent 7.7% of the overall dataset. According to De Angeli et al. (2022), the class imbalance is when the proportion of samples in a dataset belonging to one or more classes changes dramatically. Since the tweets are downloaded via keyword searching, thus, tweets with CNs can be described as less common in the dataset. Based on Guo, Yin, Dong, Yang and Zhou’s (2008) findings, the majority class tends to overwhelm

standard classifiers, causing them to overlook the minority class; this suggests that the imbalance produces unsatisfactory classification performance and that most algorithms perform poorly.

Multiple studies have proposed various remedies for the machine learning models with imbalanced datasets (Guo, Yin, Dong, Yang & Zhou, 2008) (De Angeli et al., 2022). In this circumstance, the model was trained with a class weight, which could be used to incorporate simple cost-sensitive learning; assigning larger weights to minority groups would push the model to pay more attention to them. These parameters of class weights determine how the model will be penalised if specific classes are misclassified.

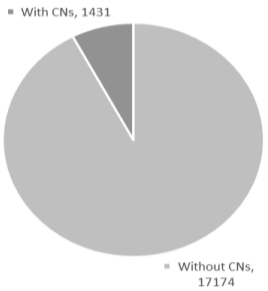

Figure 5. The distribution of the two classes.

The model used the following equation for the class weight implementation, where minority classes are assigned non-excessive, larger weights. The weight c stands for the weight for class c, while |y min| is the amount of minimum class and |y c| is the number of class c.

$$weight_c = \frac{|y_{min}|}{|y_c|}$$

4.2 The general performance of the Text CNN model

Category	Precision	Recall	F1-score
Without CNs	0.93	0.64	0.76
With CNs	0.10	0.46	0.16
Macro avg	0.51	0.55	0.46

Table 2. Performance metrics of Text CNN.

After confirming the architecture of Text CNN and weights of classes, performance metrics for the performance on the validation dataset are listed above. As De Angeli et al. (2022) indicated, accuracy and *overall* F1 score may be misleading toward the judgement on model performance since the majority classes will account for a large percentage of the score, and data about the model's performance in minority classes will be lost. It is also essential to compute macro F1 scores to *indicate* overall performance because it averages the model's performance on specific classes regardless of their frequency in the datasets. Thus, the macro F1 score is a popular assessment metric used in issues with class imbalance.

In this study, the emphasis is mainly on identifying the texts with CNs; thus, the F1 score for "With CNs" should be the main metrics to evaluate the model. However, if the model aims to identify multiple categories, the macro F1 score can better suggest the quality of outcomes. Based on table 2, around 46% of tweets with CNs in the validation set can be recognised by this model; even though the precision is only 0.1, nevertheless, after manual selection, most false positives are expected to be excluded. The setting of performance metrics will be further discussed in the discussion session.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$Overall\ F1 = 2 \times \frac{Precision \times Recall}{(Precision + Recall)}$$

$$Macro\ F1\ = \frac{1}{|C|} \times \sum_{C_i}^{C} F1(C_i)$$

Class i's F1 score is F1(C_i), and the total number of classes in the dataset is |C|; the precision and recall refer to the overall precision and recall for all categories.

4.3 Extracted innovation ideas and product categories

This session intends to illustrate the overall extracted flavours from the labelled tweets. Overall, there are 1431 tweets were labelled as containing CNs, and 1734 flavours were identified, which means that Each tweet carries an average of 1.21 flavour needs.

4.3.1 Flavour innovations

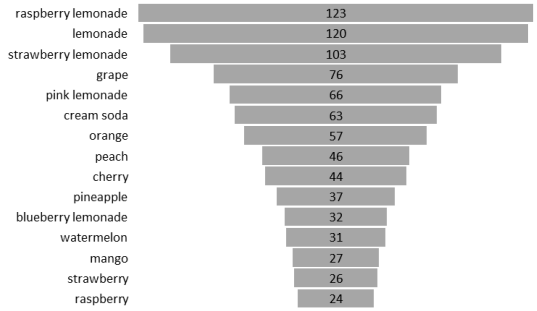

Figure 1. Occurrences of top 15 flavours

Based on the statics data of flavours, the top 3 flavours were mentioned by 346 times, while all tastes are lemonade or composite flavour of lemonade. It indicates that lemonade is the most popular flavour, while most lemonade refers to the lemon flavour. This may suggest that lemon can be a good base for a composite flavour.

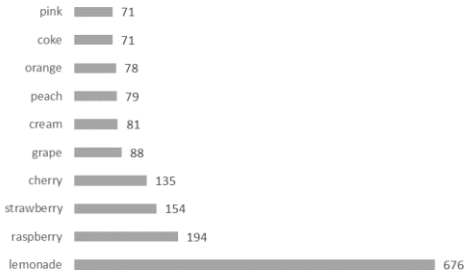

Figure 2. Occurrences of words

The distribution of words also provides insights into customers' preferred flavours. As mentioned above, lemonade not only appears as a sing flavour but also as a combination taste. Various berries are constantly paired with "lemonade" across different tweets, and the comfort zone for customers is the refreshing feeling beverages bring. Besides, grape, orange, cream soda and peach are the most frequently mentioned words. Even though these beverage flavours exist on the current market, most

beverage drinkers express strong emotions (either positive or negative) towards the same flavour from different brands.

In accordance with the above information, the tendency of flavours to focus on the fruit taste, besides the compounded flavour, are the fruits commonly appear in English-spoken countries since the language of the tweets is limited to English. However, due to the influence of Asian beverages, for instance, Korean soda and ramune Japanese soda, oriental fruits (i.e., yuzu, lichi) are also craved in some tweets. Except for the fruity flavours, other types of taste are notable, for instance, milk, matcha, lavender and nuts.

5. DISCUSSION

5.1 The Impact factors for model performance

5.1.1 Improving the test scores through adjusting parameters

As the initial step of processing input sentences, transforming a sentence matrix with different pre-trained word vectors has a critical impact on the performance of the Text CNN model. Based on the studies of Kim (2014), which differentiate Text CNN models into CNN-rand, CNN-static, CNN-non-static and CNN-multichannel, each variant model may perform well in a specific dataset. Zhang and Wallace (2015) also added that, under the circumstance of using CNN-non-static, various word vectors (Word2vec, GloVe or joint vectors) have diverse influences on model performance.

This study adopted the Glove with a CNN-static to convert the sentences; this pattern may perform poorly in this tweet dataset, especially since the processed sentences are unstandardised (as indicated above, tweets contain acronyms, shortenings, et al.). Moreover, CNN-rand was proven underfitting for this dataset, as the model considered all tweets in the validation set as tweets without CNs. CNN-non-static refers to models with fine-tuned pre-trained vectors for tasks; CNN-multichannel refers to models with two sets of word vectors which may fine-tune one set of vectors while keeping the other constant.

Furthermore, as Zhang and Wallace (2015) indicated, the regional size of filters may significantly affect the model performance. The study stressed that the combining filter region sizes should be close to the best single region size; otherwise, the combining filter size may negatively influence the test score. During modifying the single region size, the test scores suggested 5 and 7 are the optimal region sizes; thus, the combining filter region sizes are set as (3,4,5), (4,5,6) and (5,6,7). After training each plan serval times for a mean performance score, it suggests that (3,4,5) work optimally for the discussed model.

5.1.2 Improvement through optimising imbalanced dataset

In the previous section, the adverse impacts of imbalanced classes were discussed. The method adopted by this study is to add a class weight when training the model to enable more emphasis on the minority class. Roy, Cruz, Sabourin and Cavalcanti (2018), Krawczyk (2016), Guo, Yin, Dong, Yang and Zhou (2008) (De Angeli et al., 2022) and De Angeli et al. (2022) have introduced and proposed various approaches to cope with this problem.

1. Pre-processing methods (e.g., SMOTE) that resample the data space to rebalance the class distribution are used in data-level approaches. As a result, they are the most often utilised techniques for dealing with unbalanced data (Roy, Cruz, Sabourin & Cavalcanti, 2018) (Guo, Yin, Dong, Yang & Zhou, 2008) (De Angeli et al., 2022).

2. The cost-sensitive learning framework approach is between the data and algorithm level techniques, comprising cost-sensitive MLP and RBF. It assigns various costs to distinct occurrences and adjusts the learning algorithm to accept the costs (De Angeli et al., 2022). This study adopts a naive cost-sensitive method.

3. MCE methods combine an ensemble learning algorithm with one of the techniques mentioned above, which is typically a pre-processing technique (Krawczyk, 2016).

De Angeli et al. (2022) also initialled a "Class-Specialized Ensemble" method. It allows individual ensemble members to understand the essential characteristics of their allocated class group (for instance, the first model is trained for classifying class 1 to class 50, and the second one is trained for identifying class 51 to 100) during the first learning period. Then, given the whole dataset to every model for fine-tuning. Even though this method cannot be used for binary text categorisation, this method may inspire further model development, for instance, classifying the products that tweets refer to since customers prefer to advise on flavours for certain drinks.

5.2 The detected CNs and innovation process

As illustrated above, the occurrences of flavours were listed by identifying the tweets with CNs and manually extracting them. This section would like to discuss how PepsiCo could utilise this information to create new flavours for competitive advantages.

5.2.1 Flavour pairing

As figure 6 shows, the most craving flavours are fruits and lemonade, which may indicate the market demand for the compounded flavours based on lemonade. Another notable point is that the tweets with desired flavours frequently companies with the brands and types of beverages (i.e., energy drinks, Italian sodas); thus, except for tastes, PepsiCo should also concentrate on the craved flavours demands for various kinds of beverages.

Apart from flavours with high occurrences, there are also other noteworthy tastes: melon, grapefruit, dragon fruit, root beer, yuzu, guava, banana, kiwi, kombucha and ginger. Some tweets also indicated the regional limitation of some flavours (flavours that cannot find in nearby areas). Nonetheless, since most tweets do not carry geographical information, PepsiCo may collect tweets with geographical data only to identify the flavour for each region.

The Text CNN model may also help to launch the seasonal flavours. Pursuant to table 3, the flavours of the year are slightly varied. However, flavours like strawberry lemonade and pink lemonade appear in both years; PepsiCo could refer to historical data for launching limited flavours without pilot tests.

Top flavours of 2021	Top flavours of 2022
strawberry lemonade	strawberry lemonade
raspberry lemonade	orange
cream soda	pink lemonade
pink lemonade	peach
grape	cream soda
orange	cherry
lemonade	coke
cherry	mango
peach	strawberry
pineapple	pineapple

Table 1. Top flavour for the year

Through the observations of preferences from Twitter, it is possible to recognise the patterns of flavour combinations (for instance, most preferred compounded tastes contain lemonade) and discover the potential flavours (i.e., honey yuzu, matcha lemonade). The market insights powered by machine learning

algorithms also indicate beverage consumers' interest in single flavours. This method enables PepsiCo to coop with the rapidly changing market needs since Text CNN is based on instant social media content. This machine-learning-based UGC analysis may empower PepsiCo to quickly take over the undiscovered market gap.

5.1.2 Innovation process and machine learning algorithms

The previous section illustrates the observed outcomes supported by Text CNN and manual extraction; these insights suggest the possible compounded and sing flavours PepsiCo could adopt for Italian sodas, energy drinks and other beverage categorisations. Via a machine learning model that scores 0.46 at the recall for tweets with CNs, around 46% of mentioned flavours are correctly pointed out; moreover, as discussed in the 5.1, the performance of the model can be improved through multiple perspectives, which means that over 46% of user needs can be appropriately recognised with fine-tuning. As 2.2 and 2.5 indicate, the model discussed in this study may help PepsiCo boost the company's performance and increase customer value via the customer insights it brings. It suggests that PepsiCo is able to monitor the instant desires on Twitter directly of beverage lovers, which empowers the ability of the company to gain market insights.

The superior insights not only enable PepsiCo to create customer orientation flavours of beverages but also increase the efficiency of innovation processes. According to Timoshenko and Hauser (2019), as a typical VOC research suggests, professional services charges account for most of the expenditures. Around 40% of these expenditures are spent on interviewing consumers, 40%–55% on identifying and winnowing customer demands from interviews, and 5%–20% on arranging customer needs into a hierarchy and generating the final report. The 40 -55 % of costs are eliminated by UGC, while the machine-learning technique provides a 15%–22% decrease in the time spent discovering and winnowing market demands and between 46 and 52% of overall professional services prices saved. These are significant cost reductions for the company and its customers, which might help with market research for new product development.

6. LIMITATIONS AND FUTURE RESEARCH

Numerous duplicate tweets cannot be recognised since the data is tweets, considering the data filtering step. These unrecognised duplicates adopt some strategies to avoid being detected by Twitter due to a violation of the "Platform manipulation and spam policy" from the official. These spams add slight differences to each Content, and this strategy works well when the filtering step use ".unique()" to seek duplicates. In further research, it is suggested to eliminate duplicates by comparing the similarity of tweets, which can be achieved via unsupervised word vectors like SVM.

Moreover, in the data pre-processing step, the methods used aim to proceed with standard texts, while tweets' linguistic characterises have a significant variant with standard texts. Thus, this step could not remove some stop words (for instance, "ahhh"), separate terms connected by/ (strawberry/mango will be tokenised as strawberrymango) and fix typing errors (waterlemon constantly appears as a typo for watermelon). Further studies may develop a strategy for processing tweets; the variants of stop words and typing errors may be solved through a typo corpus or correlation algorithms (vector distance or Levenshtein distance).

From the perspective of model performing, to achieve a higher level of performance, the number of filers is set as 512. It increases the costs of calculation, however, lowering the number of filters will cause the problem of overfitting. This research adds a dropout layer to prevent overfitting; however, the problem may still exist due to the low dropout rate. Thus, the further model may utilise k-fold cross-validation, add L2 regulation into convolution layers, attempt various kernel initialisers for weight Initialisation and apply LSTM architecture. Besides, class weight also limits the application of the model to text classification.

De Angeli et al. (2022) indicated that adding class weights may help identify the minority class but lower the test performance of the majority class. The trade-offs in micro F1 score and macro F1 score is a hard decision, but in this case, the model focus on the performance of the minority class. Thus, the costs of adding class weight are acceptable; however, when facing multi-classification problems (i.e., identifying the beverage categories), class weighting may limit the application of Text CNN. Therefore, it suggests that further research may adopt various or joint strategies discussed above to improve versatility.

The performance of Text CNN is also influenced by the pooling size, optimiser function, vocabulary size, kernel initialiser, stride, padding method and other parameters. However, this research does not illustrate the impacts of tuning these settings; therefore, for various tasks, it is essential to observe the variability of performances and achieve the desired state by deciding the trade-off between high performance and a long training period.

From the side of flavour extraction, the significance of flavours is decided via the occurrences of terms. Nevertheless, the number of indicates cannot be treated as the degree of desire in some cases; for instance, though the flavour of "pink lemonade" is mentioned frequently, some customers also exude negative attitudes toward the flavour. Since the emphasis is on studying tweets with CNs, tweets with or partly carrying negative emotions would not be extracted, which leads to neglect of flavours aversion. Progressive combined models can be developed based on this idea; the first layer of the model may be used to distinguish tweets about beverages. Then, the second layer can be used to conduct sentiment analysis; subsequently, tf-idf-like weight calculation algorithms considering pessimistic and optimistic opinions can be employed in innovation significance analysis. Timoshenko and Hauser (2019) applied another importance analysis that uses surveys' experiential interviews to compute significance.

7. CONCLUSION

All in all, this study develops a Text CNN base tweets classifier to distinguish numerous tweets that contain specific keywords and carry certain flavour needs. In the following parts, multiple pathways for improving the model performance are discussed; moreover, based on manually labelled 1431 tweets, this research suggests various applications on the generated list of word occurrences. For instance, the available data indicates that lemonade can be regarded as a good base taste. However, the report based on extracted tweets should not be limited to occurrences analysis. The limitation partsb illustrate the limitations of the current model and report generated based on the machine learning method. Nevertheless, the existing models and data could still provide PepsiCo with valuable information on market preferences. Thus, it suggests that PepsiCo could develop machine learning-based tweets classifiers and a well-organised report based on extracted information for gaining customer insights and developing new flavours based on generated messages.

From the perspective of impacts on the innovation process, as research indicate, the machine learning-based UGC analysis could efficiently collect market insights compared with traditional research methods. Thus, adopting such a method may reduce market research costs and empower its cost-efficiency. Nonetheless, the impacts of machine learning-based UGC analysis should be further discussed in business contexts for better acknowledgement of its functioning.

8. ACKNOWLEDGEMENTS

Herewith, I would like to express my very great appreciation to dr. Dorian Proksch for offering me an excellent opportunity to experience the boundary of data science and business innovation; and for guiding me with his unparalleled patience. I would also like to extend my thanks to my second supervisor Dr. Tim G. Schweisfurth, for his supervision during the thesis. Finally, I wish to thank fellow students in the bachelor circle, Selina Noorlander, Luisa Rensing, Aaron Sandberg and Theodor Manole, for the kind communications and support during these months.

9. REFERENCE LIST

[1] Bigliardi, B., Ferraro, G., Filippelli, S., & Galati, F. (2020). Innovation models in food industry: A review of the literature. J ournal of technology management & innovation, 15(3), 97-107.
[2] Bird, S., Klein, E., & Loper, E. (2009). Natural language pro cessing with Python: analyzing text with the natural language to olkit. " O'Reilly Media, Inc."
[3] Boot, A. B., Tjong Kim Sang, E., Dijkstra, K., & Zwaan, R. A. (2019). How character limit affects language usage in tweet s. Palgrave Communications, 5(1), 1-13.
[4] Christensen, K., Nørskov, S., Frederiksen, L., & Scholderer, J. (2017). In search of new product ideas: Identifying ideas in o nline communities by machine learning and text mining. Creati vity and Innovation Management, 26(1), 17-30.
[5] Coussement, K., & Van den Poel, D. (2008). Improving cust omer complaint management by automatic email classification using linguistic style features as predictors. Decision support sy stems, 44(4), 870-882.
[6] De Angeli, K., Gao, S., Danciu, I., Durbin, E., Wu, X., & St roup, A. et al. (2022). Class imbalance in out-of-distribution dat asets: Improving the robustness of the TextCNN for the classifi cation of rare cancer types. Journal Of Biomedical Informatic s, 125, 103957. doi: 10.1016/j.jbi.2021.103957
[7] de Camargo Fiorini, P., Seles, B. M. R. P., Jabbour, C. J. C., Mariano, E. B., & de Sousa Jabbour, A. B. L. (2018). Manage ment theory and big data literature: From a review to a research agenda. International Journal of Information Management, 43, 1 12-129.
[8] Fischer, E., & Reuber, A. R. (2011). Social interaction via n ew social media:(How) can interactions on Twitter affect effect ual thinking and behavior?. Journal of business venturing, 26 (1), 1-18.
[9] Flavor as a Business Building Strategy. (2006). Retrieved 1 April 2022, from https://www.preparedfoods.com/articles/1053 50-flavor-as-a-business-building-strategy
[10] Guo, B., Zhang, C., Liu, J., & Ma, X. (2019). Improving te xt classification with weighted word embeddings via a multi-ch annel TextCNN model. Neurocomputing, 363, 366-374.
[11] Guo, X., Yin, Y., Dong, C., Yang, G., & Zhou, G. (2008). On the Class Imbalance Problem. 2008 Fourth International Co nference On Natural Computation. doi: 10.1109/icnc.2008.871
[12] Griffin, A., & Hauser, J. R. (1993). The voice of the custo mer. Marketing science, 12(1), 1-27.
[13] Jeong, B., Yoon, J., & Lee, J. M. (2019). Social media min ing for product planning: A product opportunity mining approac h based on topic modeling and sentiment analysis. International Journal of Information Management, 48, 280-290.
[14] Kaufhold, M. A., Rupp, N., Reuter, C., & Habdank, M. (20 20). Mitigating information overload in social media during con flicts and crises: design and evaluation of a cross-platform alerti ng system. Behaviour & Information Technology, 39(3), 319-3 42.
[15] Kemp, S.E. (2013). Open innovation in the Food and Beve rage Industry || Consumers as part of food and beverage industr y innovation. , (), 109–138. doi:10.1533/9780857097248.2.109
[16] Kim, Y. (2014). Convolutional Neural Networks for Sente nce Classification. Proceedings Of The 2014 Conference On E mpirical Methods In Natural Language Processing (EMNLP). d oi: 10.3115/v1/d14-1181
[17] Krawczyk, B. (2016). Learning from imbalanced data: ope n challenges and future directions. Progress In Artificial Intellig ence, 5(4), 221-232. doi: 10.1007/s13748-016-0094-0
[18] Krumm, J., Davies, N., & Narayanaswami, C. (2008). User -Generated Content. IEEE Pervasive Computing, 7(4), 10-11. d oi: 10.1109/mprv.2008.85
[19] Kühl, N., Scheurenbrand, J., & Satzger, G. (2020). Needmi ning: Identifying micro blog data containing customer needs. ar Xiv preprint arXiv:2003.05917.
[20] Kühl, N., Mühlthaler, M., & Goutier, M. (2019). Supportin g customer-oriented marketing with artificial intelligence: auto matically quantifying customer needs from social media. Electr onic Markets, 30(2), 351-367. doi: 10.1007/s12525-019-00351-0
[21] Lee, T. Y., & Bradlow, E. T. (2011). Automated marketing research using online customer reviews. Journal of marketing r esearch, 48(5), 881-894.
[22] Lies, J. (2019). Marketing intelligence and big data: Digital marketing techniques on their way to becoming social engineer ing techniques in marketing. International Journal of Interactive Multimedia & Artificial Intelligence, 5(5).
[23] Liu, Y., Jiang, C., & Zhao, H. (2019). Assessing product c ompetitive advantages from the perspective of customers by mi ning user-generated content on social media. Decision Support Systems, 123, 113079.
[24] Lomax, W., & McWilliam, G. (2001). Consumer response to line extensions: Trial and cannibalisation effects. Journal of Marketing Management, 17(3-4), 391-406.
[25] Majava, J., Nuottila, J., Haapasalo, H., & Law, K. M. (201 4). Customer needs in market-driven product development: Pro duct management and R&D standpoints. Technology and Invest ment, 2014.
[26] Massoudian, N. (2016). Twitter sentiment analysis to study association between food habit and diabetes (Doctoral dissertat ion, Wichita State University).
[27] McKinney, W., & others. (2010). Data structures for statist ical computing in python. In Proceedings of the 9th Python in S cience Conference (Vol. 445, pp. 51–56).
[28] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & De an, J. (2013). Distributed representations of words and phrases a nd their compositionality. Advances in neural information proce ssing systems, 26.
[29] NLP Tutorial - Javatpoint. Retrieved 3 April 2022, from htt ps://www.javatpoint.com/nlp3.3 Data labelling
[30] O'Shea, K., & Nash, R. (2015). An introduction to convolu tional neural networks. arXiv preprint arXiv:1511.08458.

[31] Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
[32] Prasarnphanich, P., & Wagner, C. (2009). Explaining the sustainability of digital ecosystems based on the wiki model through critical-mass theory. IEEE transactions on industrial electronics, 58(6), 2065-2072.
[33] Rahman, M. R., & Safeena, P. K. (2016). Customer Needs and Customer Satisfaction.
[34] Roy, A., Cruz, R., Sabourin, R., & Cavalcanti, G. (2018). A study on combining dynamic selection and data pre-processing for imbalance learning. Neurocomputing, 286, 179-192. doi: 10.1016/j.neucom.2018.01.060
[35] Salminen, J., Jung, S. G., & Jansen, B. J. (2021, October) Manual and Automatic Methods for User Needs Detection in Requirements Engineering: Key Concepts and Challenges. In 2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME) (pp. 1-7). IEEE.
[36] Saura, J. R., Ribeiro-Soriano, D., & Palacios-Marqués, D. (2021). From user-generated data to data-driven innovation: A research agenda to understand user privacy in digital markets. International Journal of Information Management, 60, 102331.
[37] Schaffhausen, C., & Kowalewski, T. (2015). Large-Scale Needfinding: Methods of Increasing User-Generated Needs From Large Populations. Journal Of Mechanical Design, 137(7). doi: 10.1115/1.4030161
[38] Shane, S. (Ed.). (2009). The handbook of technology and innovation management. John Wiley & Sons.
[39] Shillito, M. L. (2000). Acquiring, processing, and deploying: Voice of the customer. Crc Press.
[40] Stringfellow, A., Nie, W., & Bowen, D. E. (2004). CRM: Profiting from understanding customer needs. Business Horizons, 47(5), 45-52.
[41] Timoshenko, A., & Hauser, J. R. (2019). Identifying customer needs from user-generated Content. Marketing Science, 38 (1), 1-20.
[42] Tuorila, H. (2007). Sensory perception as a basis of food acceptance and consumption. Consumer-led food product development, 34-65.
[43] VanderPlas, J. (2016). Python data science handbook: Essential tools for working with data. “O’Reilly Media, Inc.”.
[44] Wang, Y., Mo, D. Y., & Tseng, M. M. (2018). Mapping customer needs to design parameters in the front end of product design by applying deep learning. CIRP Annals, 67(1), 145-148.
[45] Xinhua (2021). Key Laboratory of Big Data Mining and Knowledge Management, Chinese Academy of Sciences: Genki Forest becomes the most popular brand of sugar-free sparkling water Sugar-free beverage market to double in 5 years-Xinhua. (2021). Retrieved 17 March 2022, from http://www.news.CN/food/20211117/b359018321b04f31a0c51d267795883c/c.html#:~:text=%E5%85%B6%E4%B8%AD%EF%BC%8C%E5%85%83%E6%B0%94%E6%A3%AE%E6%9E%97%E5%A4%8D%E5%90%88%E5%A2%9E%E9%95%BF,%E5%B8%82%E5%9C%BA4.5%25%E5%92%8C1.8%25%E3%80%82
[46] Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners’ guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820.