Information Processing and Management 58 (2021) 102389

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Mining product innovation ideas from online reviews

Min Zhang[a], Brandon Fan[b], Ning Zhang[c], Wenjun Wang[d], Weiguo Fan[e,*]

[a] *Informatics, University of Iowa, Iowa City, IA, USA*
[b] *University of Michigan, Ann Arbor, USA*
[c] *Qingdao University, China*
[d] *College of Business, University of Arkansas at Little Rock, Little Rock, AR, USA*
[e] *Tippie College of Business, University of Iowa, Iowa City, IA, USA*

ARTICLE INFO

Keywords:
Information extraction
Deep learning
Product innovation
Online review mining
Text classification
Word embedding

ABSTRACT

The importance of online customer reviews to product innovation has been well-recognized in prior literature. Mining online reviews has received extensive attention and efforts. Most existing research on mining online reviews focus on issues such as the impact of reviews on sales, helpfulness of reviews, and customers' participation in reviews. Few research studies, however, seek to identify and extract innovation ideas for products from online reviews. This type of information is particularly important for product functionality improvement and new feature development from a manufacturer's perspective. Mining product innovation ideas allows a manufacturer to proactively review customer opinion and unlock insights about new functionality and features that the market expects, in order to gain a competitive advantage. In this paper, we propose a deep learning-based approach to identify sentences that contain innovation ideas from online reviews. Specifically, we develop a novel ensemble embedding method to generate semantic and contextual representations of the words in review sentences. The resultant representations in each sentence are then used in a long short-term memory (LSTM) model for innovation-sentence identification. Moreover, we adopt a focal loss function in our model to address the class imbalance problem. We validate our approach with a dataset of 10,000 customer reviews from Amazon. Our model achieves an AUC score of 0.91 and an F1 score of 0.89, outperforming a set of state-of-the-art baseline models in the comparison. Our approach can be extended and applied to many other information extraction tasks.

1. Introduction

In order to maintain a competitive advantage in a constantly changing environment, it is important for companies to remain on the cutting edge of innovation. An innovation in business is the process of translating an idea or invention into a good or service that creates value for which customers will pay.[1] Thus, generating new ideas or creative thoughts that meet customer needs (hereafter referred to as *innovation ideas*) is essential in the innovation process. To glean innovation ideas, companies rely on not only internal teams of professionals, but also external resources such as product users. The role of users as a source of innovation in industry was first reviewed and documented in 1978 by Eric von Hippel (von Hippel, 1978). Since then, extensive studies have provided empirical evidence to support the importance of product users in innovations (Chatterji & Fabrizio, 2012; Hienerth, Von Hippel & Berg Jensen,

* Corresponding author.
E-mail address: weiguo-fan@uiowa.edu (W. Fan).

[1] https://businessdictionary.com/

https://doi.org/10.1016/j.ipm.2020.102389
Received 15 November 2019; Received in revised form 10 September 2020; Accepted 11 September 2020

2014; Jong, 2019; Kaiserfeld, 2017; Schweisfurth, 2017; von Hippel, 2016). Classical ways for companies to obtain ideas from users include observing of a group of users for an extended period, conducting in-depth interviews, running focus groups, and working with lead users (Cooper & Edgett, 2008). However, these methods are often constrained by factors such as location, internal resources, user groups, and time. To tackle these challenges, companies developed online innovation communities to allow users to suggest, discuss, and vote on different new product and service ideas (Grimpe, Sofka, Bhargava & Chatterjee, 2017). For example, Dell launched Dell *IdeaStorm* in 2007, which enabled Dell to directly interact with its customers and collect ideas for product improvements or innovations.[2] Similar online communities can be found in many large companies such as Starbucks, P&G, XiaoMi, and Haier. Although those online communities greatly facilitates the generation of innovation ideas (Dahlander & Wallin, 2006), it might only be applicable to certain product categories that have groups of dedicated and enthusiastic supporters (Cooper & Edgett, 2008).

Online customer reviews posted on a company's website or a third-party's website, provide feedbacks of various products from a large number of users. The importance of online customer reviews to product innovation was well-recognized in prior literature (Jin, Ji & Gu, 2016; Li, Hitt & Zhang, 2011; Qi, Zhang, Jeon & Zhou, 2016; Zhan, Loh & Liu, 2009; Zhou et al., 2018). Particularly, a recent work confirmed that online reviews can help companies improve their innovations to meet users' demands (Zhang, Rao & Feng, 2018). Yet, mining innovation ideas from online reviews remains understudied. The majority of the studies in the online review literature focus on issues such as the impact of reviews on sales (Chevalier & Mayzlin, 2006; Duan, Gu & Whinston, 2008; Forman, Ghose & Wiesenfeld, 2008; Sonnier, Mcalister & Rutz, 2011), customers' participation in reviews (Chen, Fay, & Wang, 2011; Kim, Mattila & Baloglu, 2011; Lee, Park & Han, 2008), and helpfulness of online reviews (Li & Zhan, 2011; Mudambi & Schuff, 2010). Though a group of studies did work on finding product defects (Abrahams, Fan, Wang, Zhang & Jiao, 2015, 2012; Qiao, Zhang, Zhou, Wang & Fan, 2017), customers' needs or requirements (Qi et al., 2016; Timoshenko & Hauser, 2019) from online reviews, these works are different from mining innovation ideas. In addition, innovation ideas are harder to detect than defects, needs, and requirements as they are more complex and conceptual.

The goal of this study is to develop an approach that enables us to accurately detect innovation ideas from online reviews. Specifically, we aim to create an effective classifier for identifying from online reviews the sentences that contain innovation ideas (hereafter referred to as *innovation sentences*). Reviews vary on length and contain heterogenous information such as delivery, packaging, product features, user experience, and suggestions for potential customers. Therefore, detecting innovation ideas at the sentence level would be more specific and more useful. Previous studies on detecting innovation ideas from online communities, product defect and customer needs from online review mainly rely on features such as bag-of-words or the vector space model to represent the text and then use classical machine learning methods to make classifications (Christensen, Nørskov, Frederiksen, & Scholderer, 2017; Lee et al., 2013; 2018). However, bag-of-words or the vector space model cannot adequately capture the context of a word in the text, and its semantic and syntactic similarity to other words in the same text. Moreover, traditional machine learning methods are not able to learn sequential dependencies in the text.

Motivated by the recent advances in deep learning and natural language processing (NLP), we propose a novel RNN-based Ensemble Embedding (REE) method that combines three widely used word embedding methods to generate semantic and contextual representations of words in each review sentence. Taking the resultant REE vectors as input, a bidirectional long short-term memory (LSTM) model (Hochreiter & Schmidhuber, 1997) is then developed to capture the embedded information in a sentence and classify whether it contains innovation ideas or not. We demonstrate that the proposed REE-LSTM method is able to capture the semantic meaning of words and their dependencies in a more accurate manner, which ultimately helps find innovation ideas in review sentences. In addition, considering that the majority of review sentences do not contain innovation ideas, we incorporate a focal loss function in our REE-LSTM model to address the class imbalance issue. Experiments on a large Amazon review dataset demonstrate that our model significantly outperforms the baseline models.

The rest of this paper is organized as follows. We start with a brief discussion of the related work in Section 2 and elaborate our methodology in Section 3. Experimental design and performance comparison are presented in Section 4. We discuss the implications of this study in Section 5 and point out future directions in Section 6, followed by a conclusion in Section 7.

2. Related research

2.1. Online review mining

Online reviews, a form of online word-of-mouth (WOMs), contain a wealth of valuable information from consumers (Dellarocas, 2003). Many research studies have been conducted in mining online review data on analyzing the impact of reviews on sales (Chevalier & Mayzlin, 2006; Duan et al., 2008; Forman et al., 2008; Sonnier et al., 2011), customers' participation or behavior in reviews (Chen, Fay, & Wang, 2011; Kim et al., 2011; Lee et al., 2008), and helpfulness of reviews (Hong, Xu, Wang & Fan, 2017; Li & Zhan, 2011; Malik & Hussain, 2018; Mudambi & Schuff, 2010). For example, Malik and Hussain (Malik and Hussain, 2018) analyzed review content and reviewer variables which can contribute to the review helpfulness. Hong et al. (Hong et al., 2017) conducted a comprehensive meta-analysis to examine determinant factors of online review helpfulness by surveying more than thirty empirical studies.

Online customer reviews have also been found to have a great impact on customer-centered product innovation (Jin et al., 2016; Li et al., 2011; Qi et al., 2016; Zhan et al., 2009; Zhou et al., 2018). For instance, Zhang et al. (Zhang et al., 2018) showed that there

[2] https://www.delltechnologies.com/

are strong correlations between change of smart phone features and online reviews. As the importance of online customer reviews to product innovation got recognized by researchers, more studies focus on product related issues. However, studies that seek to develop advanced machine learning methods to automatically extract product innovation ideas from online reviews are rarely seen in the literature. One stream of the product related research aims to extract defect information from online reviews. For example, Abrahams et al. (Abrahams et al., 2012) used logistic regression to discover vehicle defects from online discussion forums. Abrahams et al. (Abrahams et al., 2015) proposed an integrated text analytic framework for product defects discovery from online user-generated contents (UGCs). Qiao et al. (Qiao et al., 2017) built an advanced LDA model to identify and acquire integral information about product defects from online reviews. Based on the definition found in Merriam-Webster,[3] a defect refers to a lack of something necessary for completeness, adequacy, or perfection. Thus, identifying a lack of something is not the same as providing innovation ideas or creative thoughts. Another research stream aims to develop various methods for identifying customer needs from online customer reviews which include perceived benefits and desired features of the product. Timoshenko and Hauser (Timoshenko & Hauser, 2019) extracted customer needs from Amazon reviews. They defined customer needs as a statement describing the benefits that customers are seeking, which can include both complaining defects of current features and suggesting new features. Innovation ideas, on the other hand, are the creative thoughts or suggestions on how to meet customer needs.

2.2. Word embeddings and deep learning for text mining

To transform human words meaningfully into a numerical form that can be understood by computers, word embeddings are widely used in text mining, which map words onto a real-valued vector space (Beltagy, Lo & Cohan, 2020; Devlin, Chang, Lee & Toutanova, 2019; Mikolov, Chen, Corrado & Dean, 2013; Pennington, Socher & Manning, 2014; Yang et al., 2019). Different types of word embeddings can be grouped into two categories: traditional embeddings (e.g., Word2Vec and Glove) and contextual embeddings (e.g., BERT and XLNet). Traditional word embedding methods learn a global word embedding regardless of the meaning of words in different context. One of the most popular traditional word embeddings is Word2Vec (Mikolov et al., 2013). A similar method, Global vectors for word representation (GloVe) is an unsupervised learning global log-bilinear regression model that learns vector representation for words (Pennington et al., 2014). Contextual embedding techniques, on the other hand, can learn different representations for polysemous words. BERT (Devlin et al., 2019), short for *bidirectional encoder representations from transformers*, is a bidirectional transformer-based language model trained on Wikipedia and BooksCorpus developed by Google. BERT has given state-of-the-art results on a wide variety of NLP tasks. XLNet (Yang et al., 2019) is a BERT-like novel word-embedding model that uses a generalized autoregressive pre-training method recently released by Carnegie Mellon University and Google. XLNet improves upon BERT on 20 tasks.

When modeling a sequence of words in a text, usually one of the traditional word embeddings will be used to represent words and passed through a recurrent neural network such as LSTM or GRU (Hochreiter & Schmidhuber, 1997) to model text data. Fig. 1 shows the traditional deep learning approach that utilizes a single word embedding that often follows the original Word2Vec model proposed by Mikolov et al. (Le & Mikolov, 2014) and Pennington et al. (Pennington et al., 2014). Recently, more research has attempted to use single or multiple contextual word embeddings such as BERT and XLNet (Akbik, Bergmann & Vollgraf, 2019, 2018; Fan, Fan, Smith & Garner, 2020; Li, Li, Fu, Masud & Huang, 2016). Compared with traditional embeddings, contextualized embeddings produce better representations of words by placing the context of a word as its priority. The combination of multiple different contextualized embeddings (e.g., concatenating BERT embeddings with XLNet embeddings) have shown to further improve the performance of downstream text mining tasks such as name entity recognition (Akbik et al., 2019; Li et al., 2016; Zhai et al., 2019).

Two methods (i.e., stacked and pooled) were proposed to combine multiple embeddings (Akbik et al., 2018; Zhai et al., 2019). Stacked embeddings involve concatenating multiple different word embeddings in a row format such that each word is a direct combination of various word embeddings. Pooled embeddings involve a pooling operation based on a certain function: minimum, maximum, and mean. This involves either taking the minimum, maximum value or average value across word embeddings and projecting each embedding into a common dimension. However, both methods have major deficiencies. Stacked embeddings, if combining multiple different embeddings, can drastically increase the dimensionality of the embeddings (Akbik et al., 2018). When stacking GloVe embeddings (traditionally N-by-512 where N is the number of words), BERT (N-by-1024) and XLNet (N-by-1024), the dimension increases to N-by-2560 in stacked embedding. This can lead to higher resource allocation and ultimately longer training time and difficulties in convergence. Pooling, on the other hand, suffers loss of information. By only taking the minimum, maximum, or average of the associated word embeddings, we are essentially "summarizing" features instead of "extracting" important features. Fig. 2 highlights the current methods of combining word embeddings.

3. Methodology – a new RNN-based ensemble embedding (REE) for innovation detection

We propose a novel RNN-based Ensemble Embedding (REE) method that combines three widely used word-embedding methods to generate semantic and contextual representations of words in each review sentence. Taking the resultant REE vectors as input, a bidirectional long short-term memory (LSTM) model (Hochreiter & Schmidhuber, 1997) is then developed to learn the embedded information in a sentence and make classification on whether it contains innovation ideas or not. In addition, considering that the majority of review sentences do not contain innovation ideas, we incorporate a focal loss function in our REE-LSTM model to address the class imbalance issue.

[3] https://www.merriam-webster.com

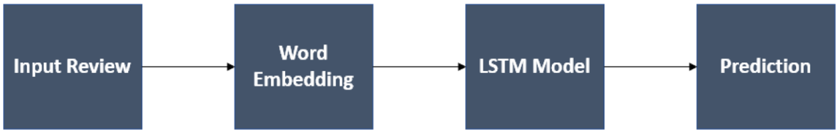


Fig. 1. Traditional deep learning approach.

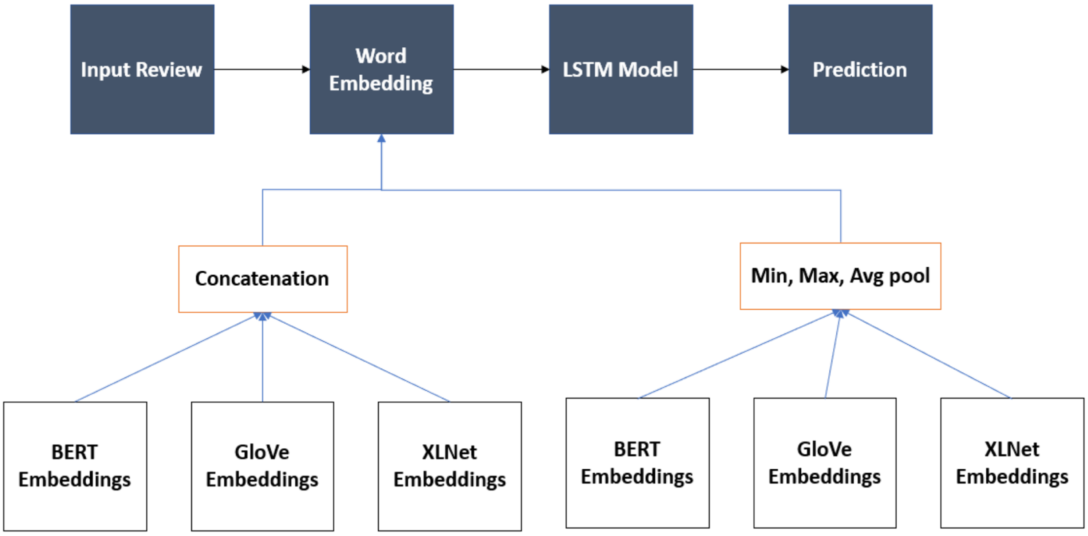


Fig. 2. Current word embedding combination methods.

3.1. RNN-based ensemble embedding (REE)

Contextualized embeddings have seen major growth within the past few years of deep learning research (Akbik et al., 2018 2019; Beltagy et al., 2020; Devlin et al., 2019; Zhai et al., 2019). Traditionally, we had the question “which word embedding produces the best performance?” when using a single embedding-based method. By combining multiple embeddings, we resolve this problem, but are faced with another question: “which combination of word embedding produces the best performance?” As we mentioned earlier, the most frequently used approaches are stacked and pooled embeddings (Akbik et al., 2018; Zhai et al., 2019). Stacked embedding concatenates multiple embeddings into a large vector (composed of the associated chosen embeddings). However, greater dimensionality leads to greater complexity and ultimately computational problems (Santos & Young, 2005). The second method utilizes pooling by summarizing multiple embedding using minimum, maximum or average (Akbik et al., 2019; Timoshenko & Hauser, 2019). Despite solving the computational complexity problem, averaging and computing simple minimums and maximum values over embeddings can lead to rich information loss and ultimately weaker performance.

Here we introduce an RNN-based ensemble method for combining word embeddings. For each word w_i that is a part of the sentence S such that $S = [w_1,\ w_2...w_n]$, we compute and concatenate GloVe ($W_{i,\,GloVe}$), XLNet ($W_{i,\,XLNet}$), and BERT ($W_{i,\,BERT}$) embeddings (Devlin et al., 2019; Pennington et al., 2014; Yang et al., 2019). We choose these three word embedding methods in our model because of their popularity and superior performance. The sentence S hence can be represented by the vector V_s, which contains a sequence of concatenated word embeddings. The sentence vector V_s is then passed through an RNN layer to produce an ensemble embedding of each concatenated word embedding in V_s. Taking a sequence of ensembled embeddings REE_S as input, a bidirectional LSTM model of 256 hidden units then computes the final representation E_s of the whole sentence, which will be passed to the sigmoid function for classification.

$$
\begin{aligned}
&S = [w_1,\ w_2...w_n]\\
&F(w_i) = \begin{bmatrix} W_{i,\ GloVe} \\ W_{i,\ XLNet} \\ W_{i,\ BERT} \end{bmatrix}\\
&V_S = [F(w_1),\ F(w_2)...F(w_n)]\\
&REE_S = RNN(V_S)\\
&E_s = \ BiLSTM(REE_S)_{t=n}
\end{aligned}
\tag{1}
$$

The proposed method trains a RNN-based ensemble embedding layer in conjunction with the bidirectional LSTM model for classification, which allows the RNN to learn what combinations of the embeddings are most important and pertinent to the

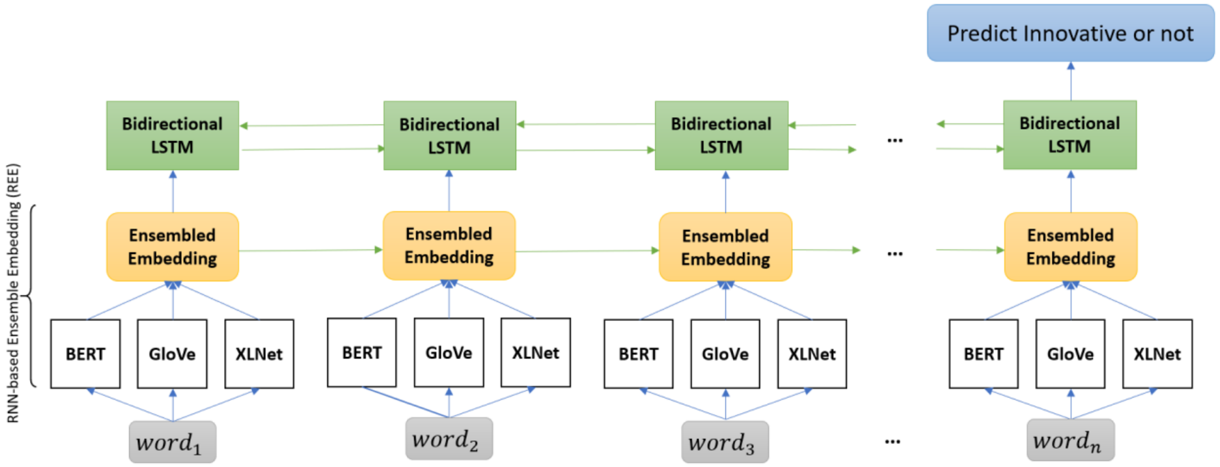


Fig. 3. RNN-based ensemble embedding model (REE).

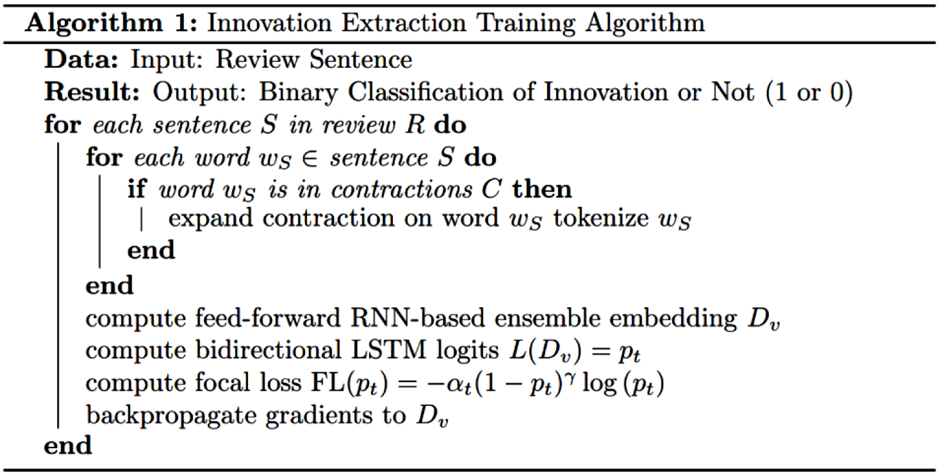
particular problem, in contrast to simply pooling or maximum operations that don't have any understanding of the data. The final state of the LSTM is used as the final embedding E_s for the model. Thus, our model reprojects the $\mathbb{R}^{N \times (300+1024+1024)}$ to an $\mathbb{R}^{512}$ dimension ($N$is the number of words in a sentence). The proposed model learns optimal features automatically while controlling the complexity of the embedding, and ultimately improves performance (as seen by the results later). A redesign of the traditional deep learning approach is shown in Fig. 3 and mathematical formulation of the REE is shown in Eq. (1). This LSTM model can be frozen or trained end-to-end with the datasets.

3.2. Loss function

Class imbalance is a common problem in machine learning, where the number of instances in one class is far less than the other class (Chen, Lin, Xiong, Luo, & Ma, 2011; Liu, Yu, Huang, & An, 2011; Vinodhini & Chandrasekaran, 2017). In our problem, only 0.21% of all sentences are labeled as innovative sentences. One strategy to handle the class imbalance problem is to either oversample the minority class or undersample the majority class. The drawback is that it may lead to either overfitting (the minority class) or underfitting (the majority class). Another strategy is to use a loss function. A loss function is basically an error function that can be used to estimate the loss of a model and then adjust the model parameters accordingly to minimize the loss. Loss functions have been widely used in various machine-learning algorithms, such as the hinge loss function, cross-entropy loss function, and mean-square-error loss function. The loss function can be also used to address the class imbalance problem. Specifically, the focal loss function has been used to compensate for class imbalance in object detection and image analysis (Lin, Goyal, Girshick, He, & Dollár, 2017). It alters the traditional binary cross entropy loss function by introducing a modulation parameter and two hyperparameters that help increase the sensitivity to misclassified instances. Utilizing focal loss, we are altering how the model learns the data by penalizing the model for predicting minority class labels incorrectly more than penalizing the model for predicting the majority class labels. There have been many approaches that implement new loss functions for deep imbalanced learning in recent studies, and many studies show that a focal loss function can solve the highly-imbalanced data problem. (Liu, Chen & Chen, 2018; Pasupa, Vatathanavaro & Tungjitnob, 2020). For example, Tran et al. (2019) implement focal loss in their model to tackle a lung nodule classification problem. The data is highly imbalanced with 1186 positive candidates and 549,879 negative candidates (0.22% positive class). By using focal loss, they boost the classification's accuracy up to 1.6%, sensitivity up to 5.8%, and specificity up to 1.3% in comparison with cross-entropy loss function (Tran et al., 2019).

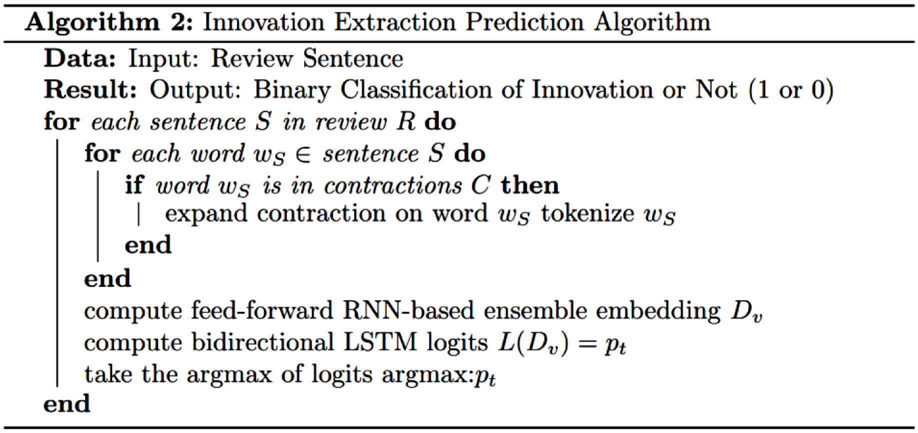
The focal loss is defined as follows. Let p be the logit probability of the binary classification of innovation or not ($y = 1$ *or* 0). We choose hyperparameters α_t and γ to optimize the binary cross entropy $-\log(p_t)$. Formally, we introduce the term $-\alpha_t(1-p_t)^\gamma$ to account for the class imbalance. This will ensure that the model gives equal weights to both classes and ultimately can distinguish between the two and reduce overfitting. Here we have an imbalance between the null class and the desired class (i.e. a binary classification). As a result, we utilize the alpha and gamma hyperparameters to "down-weight" the easy examples (i.e. the classification of the null values). Thus, the loss penalizes the model for predicting certain values with too much confidence while subsequently paying more attention to the imbalanced class.

$$p_t = \begin{cases} p, \; if\; y = 1 \\ 1-p, \; otherwise \end{cases}$$
$$FL(\mathrm{p}_t) = -\alpha_t(1-p_t)^\gamma \log(p_t) \tag{2}$$



```
Algorithm 1: Innovation Extraction Training Algorithm
Data: Input: Review Sentence
Result: Output: Binary Classification of Innovation or Not (1 or 0)
for each sentence $S$ in review $R$ do
    for each word $w_S \in$ sentence $S$ do
        if word $w_S$ is in contractions $C$ then
            expand contraction on word $w_S$ tokenize $w_S$
        end
    end
    compute feed-forward RNN-based ensemble embedding $D_v$
    compute bidirectional LSTM logits $L(D_v) = p_t$
    compute focal loss $\mathrm{FL}(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t)$
    backpropagate gradients to $D_v$
end
```

Fig. 4. Innovation extraction training algorithm.



```
Algorithm 2: Innovation Extraction Prediction Algorithm
Data: Input: Review Sentence
Result: Output: Binary Classification of Innovation or Not (1 or 0)
for each sentence $S$ in review $R$ do
    for each word $w_S \in$ sentence $S$ do
        if word $w_S$ is in contractions $C$ then
            expand contraction on word $w_S$ tokenize $w_S$
        end
    end
    compute feed-forward RNN-based ensemble embedding $D_v$
    compute bidirectional LSTM logits $L(D_v) = p_t$
    take the argmax of logits argmax:$p_t$
end
```

Fig. 5. Innovation extraction prediction algorithm.

3.3. Proposed innovation extraction algorithmic model architecture

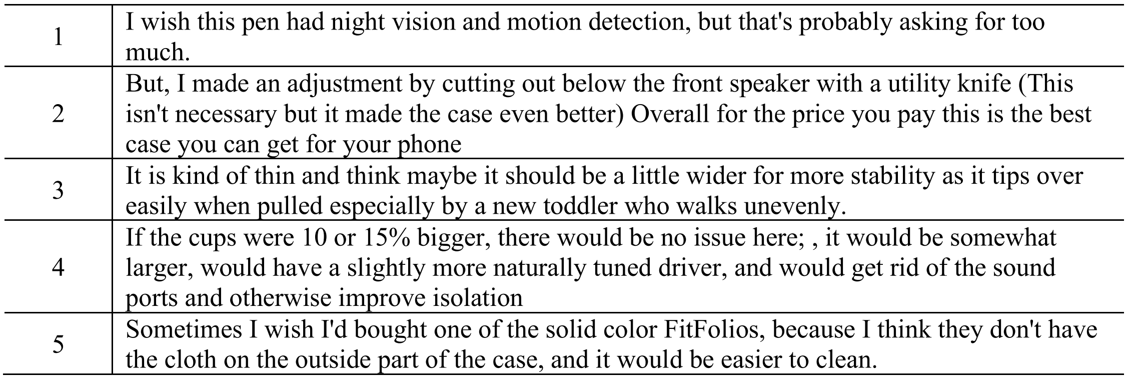
Our proposed innovation extraction model echoes our previously stated information extraction framework. Particularly, we integrate multiple different word embeddings and then passed through a text classification algorithm at the sentence-level. For training, focal loss is computed on the targeted values, ensuring that the resulting model properly distinguishes between the two classes and is not simply overfitting on one class or the other. Backpropagation is done through the ensemble embeddings to ensure that the model fine-tune the ensemble embeddings to get the most relevant information. The innovation extraction training algorithm is shown in Fig. 4 while the innovation extraction prediction algorithm is shown in Fig. 5.

4. Experiments

4.1. Data

In this study, we used a publicly available Amazon customer review dataset (He & McAuley, 2016). The dataset consisted of 10,000 randomly selected reviews from 2.8 milion review documents for 20 product categories, including Electornics, Beauty, Home and Kitchen, etc., spanning the period from May 1996 to July 2014. We filter this dataset to only contain detailed reviews and ask human labelers to go through the reviews and add innovation labels. The reviewers were tasked to decide whether a review contains sentences that relate to product innovation or not based on the definitions from Merriam-Webster.[4] and Business

[4] https://www.merriam-webster.com



1	I wish this pen had night vision and motion detection, but that's probably asking for too much.
2	But, I made an adjustment by cutting out below the front speaker with a utility knife (This isn't necessary but it made the case even better) Overall for the price you pay this is the best case you can get for your phone
3	It is kind of thin and think maybe it should be a little wider for more stability as it tips over easily when pulled especially by a new toddler who walks unevenly.
4	If the cups were 10 or 15% bigger, there would be no issue here; , it would be somewhat larger, would have a slightly more naturally tuned driver, and would get rid of the sound ports and otherwise improve isolation
5	Sometimes I wish I'd bought one of the solid color FitFolios, because I think they don't have the cloth on the outside part of the case, and it would be easier to clean.

Fig. 6. Labelled innovation sentences.

Dictionary.[5]They also highlighted sentences that were innovation-related. Each review was labelled by three labelers (kappa score of 0.87) (Fleiss, Levin & Paik, 2003). To determine the final labels, we utilized a majority vote mechanism. Based on this labeled review-level dataset, we produced a sentence-level dataset by labeling all the selected sentences as positive and non-selected sentences as negative, producing 115,489 labeled sentences. Within this sentence-level dataset, only 0.21% of all sentences are labeled as innovative sentences (a total of 243 total sentences with associated labels). Therefore, this is an extremly imbalanced dataset. Examples of labelled innovation sentences are provided in Fig. 6. Before running the baseline machine learning experiments, all sentences were preprocessed using NLTK (Loper & Bird, 2002). We lowercased each sentence, removed stop-words and punctuation, and expanded contractions. For our proposed model, no preprocessing was done to maintain the sentence structure and was ill-advised in modern transformer methods (Devlin et al., 2019; Reimers et al., 2020).

4.2. Evaluation metrics

Given the task of classifying whether a sentence is innovative or not based on a extremly imbalanced dataset, a model could achieve high accuracy by predicting only the majority class; therefore, we adopted four different metrics to evaluate the performance of a predictive model including recall, precision, F1-Score, and area under a receiver operating chracteristic curve (ROC curve). These have been frequently used as evaluation measures for imbalanced data problems (Haixiang et al., 2017). A desired classifier should have high values for each of the following evalution metrics.

- **Recall** refers to the percentage of innovative sentences correctly identified by a predictive model and is defined as $\frac{True\ Positive}{True\ Positive + False\ Negtive}$.
- **Precision** refers to the percentage of innovative sentences among all the sentences classified as innovative by a predictive model and is defined as $\frac{True\ Positive}{True\ Positive + False\ Positive}$.
- **F1-Score** is the harmonic mean of the precision and recall and is defined as $2 \cdot \frac{Precision\ \cdot Recall}{Precision + Recall}$.
- **AUC** refers the area under a receiver operating chracteristic curve (ROC curve). The ROC curve shows the true positive rate against false positive rate at various discrimination threshold. AUC is one of the widely used evaluation metrics for assessing a classifier's performance on a imbalanced dataset. It indicates how much a classifier is capable of distinguishing between classes.

4.3. Experimental design and setup

4.3.1. Experimental design

To demonstrate the superiority of our proposed method, we conduct three experiments. In the first experiment, we compare the performance of three traditional machine learning methods, as utilized in extant research (Abrahams et al., 2012; Liu et al., 2011; Qiao et al., 2017) with different tyes of inputs, with that of a deep learning method. Then we examine the predictive power of different embedding combinations on the three embedding methods: GloVe, XLNet, and BERT in the second experiment. In the last experiment, we assess the focal loss function on the proposed deep learning method.

4.3.2. Experimental data setup

To compare the average performance of predictive models, we split the raw data into 70% training and 30% test data multiple times. For each split, due to the extremely imbalanced classes, the number of non-innovation instances we selected were 10 times of innovative instances in each split, resulting in a class imbalance of 10:1. All predictive models were trained on the current split of training data and tested on the current split of test data. For instance, if we split raw data into training and test data k times, we would

[5] https://businessdictionary.com/

Table 1
Comparisons between traditional machine learning with two input types and a deep learning approach.

	Model	AUC	Precision	Recall	F1
Vector Space Model	Naïve Bayes	0.65	0.13	**0.81**	0.23
	Logistic Regression (L1)	0.76	**0.78**	0.52	0.72
	SVM	0.76	0.66	0.54	0.63
Word Embeddings	Naïve Bayes	0.61	0.28	0.31	0.29
	Logistic Regression (L1)	0.69	0.49	0.51	0.50
	SVM	0.73	0.48	0.58	0.53
	LSTM	**0.80**	0.72	0.80	**0.76**

have *k* different sets of measurement (recall, precision, F1-score, and AUC) on a predictive model. Given these *k* sets of measurement, we report the average performance on recall, precision, F1-score, and AUC for each predictive model. During the phase of training, we used cross-validation to select the best hyperparameters for each predictive model.

4.4. Experimental results

4.4.1. Traditional machine learning vs. deep learning

As aforementioned, we first compare the performance of three traditional machine learning methods with that of the deep learning method on the task of classifying whether a sentence is innovative or not. We implement Naïve Bayes, Logistic Regression with L1 regularization, Support Vector Machine (SVM) as machine learning baselines. These three methods were chosen because they were commonly used in binary text classification tasks (Abrahams et al., 2015; Gruss, Abrahams, Fan & Wang, 2018). The traditional methods expect structured data as input. However, each sentence may have a different length. Because of this, we cannot use word embeddings directly as input. To address this issue, a term frequency-inverse document frequency (TF-IDF) matrix with principal component analysis (PCA) and Word2Vec word embedding for text classification are commonly used (Buntine & Jakulin, 2004; Ma & Zhang, 2015; Schmidt, 2019; Zhang & Wallace, 2015). Therefore, we used two different ways to run the benchmark experiments. In the first approach, we used the traditional vector space model in which we calculated the term frequency-inverse document frequency (TF-IDF) matrix for the 1500 most occurring words to represent each sentence and performed principal component analysis (PCA) to reduce the input dimension to a fixed size. In the second approach, we created a fix-sized embedding of each sentence by computing a simple average of word embeddings corresponding to the words in a sentence. We used Google's pretrained Word2Vec method to get an embedding of each word in a sentence. If there are N sentences in a training dataset and each sentence has an embedding of size d, then we have a $N \times d$ matrix, which is the input of the three traditional machine learning methods. For the deep learning approach, we constructed a LSTM model with an embedding layer. Given that the LSTM can have a variable length input, our LSTM model takes a sequence of words in a sentence as the input.

We present the performance comparisons on both input approaches on the innovative sentence classification task in Table 1. It can be observed from Table 1 that LSTM overall achieved better performance compared with Naïve Bayes, Logistic Regression (L1) and SVM. The capability of learning from variable length input helps LSTM to capture richer information at the word-level.

4.4.2. Combinations of different embeddings

Given its superior performance over the three traditional machine learning methods, an LSTM is the desired approach. The LSTM built in the previous experiment takes a sequence of words in a sentence as input. In practice, there are many different ways to encode a word. In this study, we took advantage of the three well-known and publicly avaliable pretrained embedding methods: GloVe, XLNet, and BERT. Instead of examining the effectiveness of each embedding method, we compared different combinations of them in this experiment.

We first compared the performance difference on incorporating different numbers of embedding methods using our proposed RNN-based ensemble embedding. The comparisons are presented in Table 2. As shown in Table 2, the LSTM model incorporating all three embedding methods achieved better performance than combining any two of them.

Besides comparing the performance difference on incorporating different numbers of embedding methods, we also wanted to examine various ways to combine them. As identified in related work, the two major ways to obtain multi-contexture mixed embedding are stacking and pooling. Therefore, we compared stacking and pooling mechanisms with the proposed RNN-based ensemble embedding in Table 3. Note that we only compared them in the case of incorporating all three embedding methods since we already

Table 2
Comparisons on incorporating different embedding methods.

Model	AUC	Precision	Recall	F1
GloVe + BERT	0.82	0.77	0.83	0.80
GloVe + XLNet	0.79	0.79	0.83	0.81
BERT + XLNet	0.83	0.84	0.86	0.85
GloVe + BERT + XLNet	**0.88**	**0.87**	**0.89**	**0.88**

Table 3.
Comparisons between GPT-2, stacked, pooled, and proposed REE when incorporating three embeddings.

Model	AUC	Precision	Recall	F1
GPT-2	0.87	0.86	0.88	0.86
GloVe + BERT + XLNet (stacked)	0.83	0.80	0.82	0.81
GloVe + BERT + XLNet (pooled)	0.86	0.83	0.87	0.85
GloVe + BERT + XLNet (REE)	**0.88**	**0.87**	**0.89**	**0.88**

observed that combining all three embedding methods can produce the best performance. As demonstrated in Table 3, incorporating three embeddings using proposed REE in LSTM outperformed the other two ways of obtaining multi-contexture embedding. We also compare our proposed REE in LSTM with the GPT-2 model which is known to produce state of the art results through an unsupervised learning method (Radford et al., 2019).

4.4.3. Performance of the focal loss function

In the last experiment, we assessed the performance of adopting focal loss function in the proposed REE-LSTM method. In Table 4, we presented the comparison between REE-LSTM with and without using focal loss function. From Table 4, it can be observed that adopting focal loss can help boost the performance on AUC, precision, and F1 score. Clearly, the REE-LSTM with focal loss model can better distinguish between innovative and non-innovative classes (after hyperparameter tuning, an alpha value of 0.8 was chosen with a gamma of 2).

5. Discussion

In the product innovation literature, companies usually use classical methods or online communities to collect innovation ideas from customers. However, both classical methods and online communities have limitations. Classical methods are constrained by factors such as location, internal resources, user groups, and time. Online communities might only be applicable to certain products that have groups of dedicated and enthusiastic supporters (Cooper & Edgett, 2008). Although online review data has been recognized as an important source of obtaining innovation ideas from customers (Jin et al., 2016; Li et al., 2011; Qi et al., 2016; Zhan et al., 2009), studies that seek to develop methods to automatically extract product innovation ideas from online reviews are rarely seen in the literature. Prior studies on detecting innovation ideas from online communities, product defects and customer needs from online reviews mainly rely on features such as bag-of-words or the vector space model to represent the text and then use classical machine learning methods to make classifications (Christensen, Nørskov, Frederiksen, & Scholderer, 2017; Lee et al., 2013, 2018). However, features using bag-of-words or the vector space model cannot adequately capture the context of a word in the text, and its semantic and syntactic similarity to other words in the same text. Moreover, classical machine learning methods are not able to learn sequential dependencies in the text.

Our study contributes to the production innovation literature and information extraction literature by providing an advanced approach that can automatically identify sentences that contain innovation ideas from online customer reviews. We develop an ensemble method that produces the optimal combinations of a set of widely used word embeddings, in order to generate semantic and contextual representations of the words in review sentences. The resultant representations in each sentence are then used in a long short-term memory (LSTM) model for innovation-sentence identification. As majority of the sentences do not contain innovation ideas, we adopt a focal loss function in our model to address the class imbalance problem. To validate our approach, we conducted experiments on comparing the performance of traditional machine learning methods with that of a deep learning method, examining the predictive power of different embedding combinations, and assessing the performance of focal loss function on our proposed method. Results show that adopting focal loss function in REE-LSTM achieved the best performance with an AUC score of 0.91 and an F1 score of 0.89 on the dataset of 10,000 customer reviews from Amazon.

Besides detecting innovation sentences from online customer reviews, the proposed method can also be extended and applied to other tasks from different fields such as healthcare, social science, and risk management. For instance, the proposed method could help companies assess risk associated with their investment by identifying sentences containing risk signals from a target company's 10 K statements. In addition to potential future business applications, this study also confirms that the focal loss, often used in image analysis, can also greatly facilitate the learning of a model on a highly imbalanced text data set. Future studies can also use the focal loss function as an alternative in performing similar kinds of information extraction studies.

Table 4.
Comparisons between REE-LSTM with and without using focal loss function.

Model	AUC	Precision	Recall	F1
GloVe + BERT + XLNet (REE)	0.88	0.87	0.89	0.88
GloVe + BERT + XLNet (REE) With Focal Loss	**0.91**	**0.95**	0.85	**0.89**

6. Limitations and future research

Like any other study, our work also has limitations. Though our labeled data reached about 10,000 reviews and 115,000 sentences, there is still a major class imbalance issue. More data collection could go into finding innovation-based reviews to improve the model. In the framework, we did not perform multiple dimension re-projections. Instead we opt for traditional $\mathbb{R}^{512}$ found in most of deep learning literature. An ablation study should be done to determine the most optimal dimension to be used for the re-projection. The LSTM model trained in our paper did not utilize attention mechanisms (Vaswani et al., 2017); however, research has shown the major merit and improvements in various NLP tasks when utilizing attention. Future research should investigate the combination of the REE embeddings with attention mechanisms to see if results are improved and if model interpretability can be determined, pinpointing key features related to innovation determination. Lastly, our baselines are by no means exhaustive. We have tried to include the ones that are representative and state-of-the-art.

Though we achieve superior results compared to current state-of-the-art algorithms, there are still many avenues for further research: (1) An analysis of innovation extraction for overseas markets could be considered, and multilingual models should be created. However, current datasets are limited and in need of greater labeling for such a multilingual model to be implemented. (2) Investigation might be necessary on labeling a larger set of data that not only indicate innovation sentences but also defect sentences (i.e. product recommendations that need to be fixed by the manufacturer). (3) Attempting to analyze the implementation of ensemble-based approaches by combining machine-learned embedding features with human-designed features. (4) Exploration of using possible semi-supervised learning such as ladder networks (Rasmus, Valpola, Honkala, Berglund, & Raiko, 2015) and virtual adversarial training (Miyato, Maeda, Koyama & Ishii, 2019) to deal with lack of labelled data. (5) Future research can investigate refining the model architecture to see whether it can further improve the performance.

7. Conclusion

In this paper, we propose an effective deep learning model, REE-LSTM, for mining product innovation ideas from online reviews. Our model differs from existing studies in several aspects. First, we provide a new perspective of product innovation on using online product reviews. We focus on extracting sentences that contain innovation/improvement ideas for product development. Second, we develop a novel RNN-based ensemble embedding (REE) method. Our model incorporates three widely used word embeddings into an ensembled embedding that can better captures the semantic information embedded in the review sentences. Third, the REE-based ensemble embedding is naturally integrated with LSTM units to build a classifier for innovation sentence identification. Finally, a focal loss function is used in our REE-LSTM model to handle the highly imbalanced classes. Experiments on Amazon reviews validate our model and demonstrate its superior performance among a set of state-of-the-art models. Our approach provides significant insights and implications for information-extraction and text classification research and practice.

Author statement

Min Zhang: data curation, conceptualization, experimental design and implementation, analysis, manuscript writing
Brandon Fan, conceptualization, experimental design and implementation, manuscript writing
Ning Zhang, data annotation
Wenjun Wang, data annotation, paper editing
Weiguo Fan, conceptualization, project administration, paper outline and editing

References

Abrahams, A. S., Fan, W., Wang, G. A., Zhang, Z., & Jiao, J. (2015). An integrated text analytic framework for product defect discovery. *Production and Operations Management, 24*(6), 975–990. https://doi.org/10.1111/poms.12303.

Abrahams, A. S., Jiao, J., Wang, G. A., & Fan, W. (2012). Vehicle defect discovery from social media. *Decision Support Systems, 54*(1), 87–97. https://doi.org/10.1016/j.dss.2012.04.005.

Akbik, A., Bergmann, T., & Vollgraf, R. (2019). Pooled contextualized embeddings for named entity recognition. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference. 1. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* (pp. 724–728). . https://doi.org/10.18653/v1/n19-1078.

Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 1638–1649). . Retrieved from https://github.com/zalandoresearch/flair.

Beltagy, I., Lo, K., & Cohan, A. (2020). SCIBERT: A pretrained language model for scientific text. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference* (pp. 3615–3620). . Retrieved from http://arxiv.org/abs/1903.10676.

Buntine, W., & Jakulin, A. (2004). Applying discrete PCA in data analysis. *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence (UAI2004)* (pp. 59–66). . Retrieved from http://eprints.pascal-network.org/archive/00000143/.

Chatterji, A. K., & Fabrizio, K. (2012). How do product users influence corporate invention? *Organization Science, 23*(4), 907–1211. https://doi.org/10.1287/orsc.1110.0675.

Chen, Enhong, Lin, Yanggang, Xiong, Hui, Luo, Qiming, & Ma, Haiping (2011). Exploiting probabilistic topic models to improve text categorization under class imbalance. *Information Processing and Management, 47*(2), 202–214.

Chen, Y., Fay, S., & Wang, Q. (2011). The role of marketing in social media: How online consumer reviews evolve. *Journal of Interactive Marketing, 25*(2), 85–94. https://doi.org/10.1016/j.intmar.2011.01.003.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research, 43*(3), 345–354. https://doi.org/10.1509/jmkr.43.3.345.

Christensen, K., Nørskov, S., Frederiksen, L., & Scholderer, J. (2017). In search of new product ideas: Identifying ideas in online communities by machine learning and text mining. *Creativity and Innovation Management, 26*(1), 17–30. https://doi.org/10.1111/caim.12202.

Cooper, R., & Edgett, S. (2008). Ideation for product innovation: What are the best methods? *PDMA Visions Magazine, 32*, 12–17.

Dahlander, L., & Wallin, M. W. (2006). A man on the inside: Unlocking communities as complementary assets. *Research Policy, 35*(8), 1243–1259. https://doi.org/10.1016/j.respol.2006.09.011.

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science, 49*(10), 1275–1444. https://doi.org/10.1287/mnsc.49.10.1407.17308 Retrieved from.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference. 1. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference* (pp. 4171–4186). . Retrieved from http://arxiv.org/abs/1810.04805.

Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter? - An empirical investigation of panel data. *Decision Support Systems*. https://doi.org/10.1016/j.dss.2008.04.001.

Fan, B., Fan, W., Smith, C., & Garner, H. (2020). Adverse drug event detection and extraction from open data: A deep learning approach. *Information Processing and Management, 57*(1), Article 102131. https://doi.org/10.1016/j.ipm.2019.102131.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). Statistical methods for rates and proportions. *Statistical Methods for Rates and Proportions*. https://doi.org/10.1002/0471445428.

Forman, C., Ghose, A., & Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research*. https://doi.org/10.1287/isre.1080.0193.

Grimpe, C., Sofka, W., Bhargava, M., & Chatterjee, R. (2017). R&D, marketing innovation, and new product performance: A mixed methods study. *Journal of Product Innovation Management, 34*(3), 360–383. https://doi.org/10.1111/jpim.12366.

Gruss, R., Abrahams, A. S., Fan, W., & Wang, G. A. (2018). By the numbers: The magic of numerical intelligence in text analytic systems. *Decision Support Systems, 113*(August), 86–98. https://doi.org/10.1016/j.dss.2018.07.004.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications, 73*, 220–239. https://doi.org/10.1016/j.eswa.2016.12.035.

He, R., & McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. *25th International World Wide Web Conference, WWW 2016* (pp. 507–517). . https://doi.org/10.1145/2872427.2883037.

Hienerth, C., Von Hippel, E., & Berg Jensen, M. (2014). User community vs. producer innovation development efficiency: A first empirical study. *Research Policy*. https://doi.org/10.1016/j.respol.2013.07.010.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Hong, H., Xu, D., Wang, G. A., & Fan, W. (2017). Understanding the determinants of online review helpfulness: A meta-analytic investigation. *Decision Support Systems, 102*, 1–11. https://doi.org/10.1016/j.dss.2017.06.007.

Jin, J., Ji, P., & Gu, R. (2016). Identifying comparative customer requirements from product online reviews for competitor analysis. *Engineering Applications of Artificial Intelligence*. https://doi.org/10.1016/j.engappai.2015.12.005.

Jong, J. D. (2019). The empirical scope of user innovation. *Revolutionizing Innovation*. https://doi.org/10.7551/mitpress/9439.003.0007.

Kaiserfeld, T. (2017). Revolutionizing innovation: users, communities, and open innovation ed. by Dietmar Harhoff and Karim R. Lakhani. *Technology and Culture*. https://doi.org/10.1353/tech.2017.0069.

Kim, E. E. K., Mattila, A. S., & Baloglu, S. (2011). Effects of gender and expertise on consumers' motivation to read online hotel reviews. *Cornell Hospitality Quarterly*. https://doi.org/10.1177/1938965510394357.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. *31st International Conference on Machine Learning, ICML 2014. 4. 31st International Conference on Machine Learning, ICML 2014* (pp. 2931–2939).

Lee, H., Choi, K., Yoo, D., Suh, Y., He, G., & Lee, S. (2013). The more the worse? Mining valuable ideas with sentiment analysis for idea recommendation. *PACIS 2013 Proceedings*.

Lee, H., Choi, K., Yoo, D., Suh, Y., Lee, S., & He, G. (2018). Recommending valuable ideas in an open innovation community: A text mining approach to information overload problem. *Industrial Management and Data Systems*. https://doi.org/10.1108/IMDS-02-2017-0044.

Lee, J., Park, D. H., & Han, I. (2008). The effect of negative online consumer reviews on product attitude: An information processing view. *Electronic Commerce Research and Applications*. https://doi.org/10.1016/j.elerap.2007.05.004.

Li, Jianqiang, Li, J., Fu, X., Masud, M. A., & Huang, J. Z. (2016). Learning distributed word representation with multi-contextual mixed embedding. *Knowledge-Based Systems, 106*, 220–230. https://doi.org/10.1016/j.knosys.2016.05.045.

Li, Jin, & Zhan, L. (2011). Online persuasion: How the written word drives WOM. *Journal of Advertising Research*. https://doi.org/10.2501/jar-51-1-239-257.

Li, X., Hitt, L. M., & Zhang, Z. J. (2011). Product reviews and competition in markets for repeat purchase products. *Journal of Management Information Systems*. https://doi.org/10.2753/MIS0742-1222270401.

Lin, Y., T, Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal Loss for Dense Object Detection. *2017 IEEE International Conference on Computer Vision (ICCV)*, 2999–3007.

Liu, W., Chen, L., & Chen, Y. (2018). Age classification using convolutional neural networks with the multi-class focal loss. *IOP Conference Series: Materials Science and Engineering. 428*https://doi.org/10.1088/1757-899X/428/1/012043.

Liu, Y., Yu, X., Huang, J. X., & An, A. (2011). Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. *Information Processing and Management, 47*(4), 617–631. https://doi.org/10.1016/j.ipm.2010.11.007.

Loper, E., & Bird, S. (2002). *NLTK: The natural language toolkit*. 63–70. Retrieved fromhttp://arxiv.org/abs/cs/0205028.

Ma, L., & Zhang, Y. (2015). Using Word2Vec to process big text data. *2015 IEEE International Conference on Big Data* (pp. 2895–2897). . https://doi.org/10.1109/BigData.2015.7364114.

Malik, M. S. I., & Hussain, A. (2018). An analysis of review content and reviewer variables that contribute to review helpfulness. *Information Processing and Management, 54*(1), 88–104. https://doi.org/10.1016/j.ipm.2017.09.004.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). E*fficient estimation of word representations in vector space*.1–12. Retrieved fromhttp://arxiv.org/abs/1301.3781.

Miyato, T., Maeda, S. I., Koyama, M., & Ishii, S. (2019). Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(8), 1979–1993. https://doi.org/10.1109/TPAMI.2018.2858821.

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on amazon.com. *MIS Quarterly, 34*, 185–200. https://doi.org/10.2307/20721420 10.2307/20721420.

Pasupa, K., Vatathanavaro, S., & Tungjitnob, S. (2020). Convolutional neural networks based focal loss for class imbalance problem: A case study of canine red blood cells morphology classification. *Journal of Ambient Intelligence and Humanized Computing*. https://doi.org/10.1007/s12652-020-01773-x.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (pp. 1532–1543). . https://doi.org/10.3115/v1/d14-1162.

Qi, J., Zhang, Z., Jeon, S., & Zhou, Y. (2016). Mining customer requirements from online reviews: A product improvement perspective. *Information and Management*. https://doi.org/10.1016/j.im.2016.06.002.

Qiao, Z., Zhang, X., Zhou, M., Wang, G. A., & Fan, W. (2017). A domain oriented LDA model for mining product defects from online customer reviews. *Proceedings of the 50th Hawaii International Conference on System Sciences (2017)* (pp. 1821–1830). . https://doi.org/10.24251/hicss.2017.222.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). (GPT-2) Language models are unsupervised multitask learners. *OpenAI Blog, 1*(8), 9.

Rasmus, A., Valpola, H., Honkala, M., Berglund, M., & Raiko, T. (2015). Semi-supervised learning with Ladder networks. *Advances in Neural Information Processing Systems,* 3546–3554.

Reimers, N., Schiller, B., Beck, T., Daxenberger, J., Stab, C., & Gurevych, I. (2020). Classification and clustering of arguments with contextualized word embeddings. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference* (pp. 567–578). . https://doi.org/10.18653/v1/p19-1054.

Santos, I. M., & Young, A. W. (2005). Exploring the perception of social characteristics in faces using the isolation effect. *Visual Cognition, 12*(1), 213–247. https://doi.org/10.1080/13506280444000102.

Schmidt, C.W. (.2019). *Improving a tf-idf weighted document vector embedding*. Retrieved fromhttp://arxiv.org/abs/1902.09875.

Schweisfurth, T. G. (2017). Comparing internal and external lead users as sources of innovation. *Research Policy*. https://doi.org/10.1016/j.respol.2016.11.002.

Sonnier, G. P., Mcalister, L., & Rutz, O. J. (2011). A dynamic model of the effect of online communications on firm sales. *Marketing Science*. https://doi.org/10.1287/mksc.1110.0642.

Timoshenko, A., & Hauser, J. R. (2019). Identifying customer needs from user-generated content. *Marketing Science, 38*(1), 1–20. https://doi.org/10.1287/mksc.2018.1123.

Tran, G. S., Nghiem, T. P., Nguyen, V. T., Luong, C. M., Burie, J. C., & Levin-Schwartz, Y. (2019). Improving accuracy of lung nodule classification using deep learning with focal loss. *Journal of Healthcare Engineering, 2019*. https://doi.org/10.1155/2019/5156416.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., & Gomez, A. N. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 2017-Decem(Nips),* 5999–6009.

Vinodhini, G., & Chandrasekaran, R. M. (2017). A sampling based sentiment mining approach for e-commerce applications. *Information Processing & Management, 53*(1), 223–236.

von Hippel, E. (1978). Successful industrial products from customer ideas. *Journal of Marketing*. https://doi.org/10.2307/1250327.

von Hippel, E. (2016). *Free innovation*. Cambridge, MA: The MIT PressArticle 9780262035217.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). *XLNet: Generalized Autoregressive Pretraining for Language Understanding,* 1–18. Retrieved from http://arxiv.org/abs/1906.08237.

Zhai, Z., Nguyen, D. Q., Akhondi, S., Thorne, C., Druckenbrodt, C., Cohn, T., et al. (2019). Improving chemical named entity recognition in patents with contextualized word embeddings. *Proceedings of the 18th BioNLP Workshop and Shared Task* (pp. 328–338). . https://doi.org/10.18653/v1/w19-5035.

Zhan, J., Loh, H. T., & Liu, Y. (2009). Gather customer concerns from online product reviews - A text summarization approach. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2007.12.039.

Zhang, H., Rao, H., & Feng, J. (2018). Product innovation based on online review data mining: A case study of Huawei phones. *Electronic Commerce Research, 18*(1), 3–22. https://doi.org/10.1007/s10660-017-9279-2.

Zhang, Y., & Wallace, B. (2015). *A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification*. Retrieved fromhttp://arxiv.org/abs/1510.03820.

Zhou, S., Qiao, Z., Du, Q., Wang, G. A., Fan, W., & Yan, X. (2018). Measuring customer agility from online reviews using big data text analytics. *Journal of Management Information Systems, 35*(2), 510–539. https://doi.org/10.1080/07421222.2018.1451956.