Mining Purchase Intent in Twitter

Rejwanul Haque[1], Arvind Ramadurai[2], Mohammed Hasanuzzaman[1], and Andy Way[1]

[1] ADAPT Centre, School of Computing, Dublin City University, Dublin, Ireland
{firstname.lastname}@adaptcentre.ie
[2] School of Computing, Dublin City University, Dublin, Ireland
arvind.ramadurai2@mail.dcu.ie

Abstract. Most social media platforms allow users to freely express their beliefs, opinions, thoughts, and intents. Twitter is one of the most popular social media platforms where users' post their intent to purchase. A purchase intent can be defined as measurement of the probability that a consumer will purchase a product or service in future. Identification of purchase intent in Twitter sphere is of utmost interest as it is one of the most long-standing and widely used measures in marketing research. In this paper, we present a supervised learning strategy to identify users' purchase intent from the language they use in Twitter. Recurrent Neural Networks (RNNs), in particular with Long Short-Term Memory (LSTM) hidden units, are powerful and increasingly popular models for text classification. They effectively encode sequences with varying length and capture long range dependencies. We present the first study to apply LSTM for purchase intent identification task. We train the LSTM network on semi-automatically created dataset. Our model achieves reasonable classification accuracy ($F_1 = 83\%$) over a gold-standard dataset. Further, we demonstrate the efficacy of the LSTM network by comparing its performance with different classifiers.

Keywords: Social Media · Purchase Intent · Mining · User Generated Content.

1 Introduction

Sharing personal thoughts, beliefs, opinions, and intents on the internet (especially on social media platforms) has become an essential part of life for millions of users all around the world. Twitter,[3] one of the popular social media platforms, where users put forth their intent to purchase products or services and looks for suggestions that could assist them. To exemplify, 'I wanna buy an iPhone this week!' indicates the user's intent for buying an Apple iPhone soon. Essentially, identification and classification of such user generated contents (UGC) have two-fold benefits: (a) commercial companies could exploit this to build their marketing tool/strategy, and (b) it could benefit social media users with the suggestions of the products or services that they want to purchase.

Gupta et al. in their study [7] investigated the relationship between users' purchase intent from their social media forums such as Quora[4] and Yahoo! Answers.[5] They

[3] https://twitter.com/
[4] www.quora.com
[5] www.answers.yahoo.com

mainly carried out text analysis (e.g. extracting features, such as purchase action words, using the dependency structure of sentences) to detect purchase intent from UGC. In another study [18], the authors investigated the problem of identifying purchase intent. In particular, they proposed a graph-based learning approach to identify intent tweets and classify them into six categories, namely 'Food & Drink', 'Travel', 'Career & Education', 'Goods & Services', 'Event & Activities' and 'Trifle'. For this, they retrieved tweets with a bootstrap method, with using a list of seed intent-indicators (e.g. 'want to'), and manually created training examples from the collected tweets. There is a potential problem in their data set since it was created based on a handful of keywords (i.e. indent-indicators). In reality, there could have many lexical variations of an intent-indicator. For example, any of these following intent-indicators can take the place of 'want to': 'like to', 'wish to', and 'need to'. Tweets often include misspelled short or long words depending on user's emotions, thoughts and state of mind. For example, when a user is really excited to buy a car soon, his purchase intent tweet can be 'I liiiiiiiiike to buy the SUV this month!!!' that includes an intent-indicator 'liiiiiiiiike to' which has a misspelled long word, 'liiiiiiiiike'.

In this work, in order to capture new tweets that are good paradigms of purchase intentions, we adopted a seed intent-indicators expansion strategy using a query expansion technique [14]. This technique has essentially helped to increase the coverage of keywords in our training data. We manually create a labeled training data with the tweets that were extracted using a python API, given the expanded seed list of the intent-indicators. In order to identify users' purchase intention in tweets, we present a RNN model [17, 20] with LSTM units [8] (cf. Section 6). To summarize, our main contributions in this paper are as follows:

1. We are the first to apply the deep learning techniques for the users' purchase intent identification task in social media platform.
2. We create a gold-standard training data set, which, in practice, can be viewed as an ideal data set for the users' purchase intent identification task in Twitter.

The remainder of the paper is organised as follows. In Section 2, we discuss related work. Section 3 presents an existing training data that was previously used in the purchase intent identification task. In Section 4, we detail how we created training data for our experiments. In Section 5, we present our experimental methodology. Section 6 presents our experimental set-up, results and analysis, while Section 7 concludes, and provides avenues for further work.

2 Related Work

Identifying wishes from texts [16, 6] is apparently a new arena in natural language processing (NLP). Notably, [16] focus on identifying wishes from product reviews or customer surveys, e.g. a desire to buy a product. They describe linguistic rules that can help detect these wishes from text. In general, the rule-based methods for identifying wishes from text proved to be effective. However, the creation of rules is a time-consuming task, and theirs coverage is not satisfactory. Detection of users' purchase intent in social media platform is close to the task of identifying wishes in product reviews or customer surveys.

In information retrieval (IR), query intent can broadly be classified into two categories: query type [10, 3] and user type [2, 13, 9]. The focus on this paper is to identify and classify tweets that explicitly express users' purchase intents. In this sense, this work can be kept under of the first category. To the best of our knowledge, the most relevant works to ours come from [7, 18]. In fact, to a certain extent, our proposed methods can be viewed as the extension of [18]. [7] investigated the problem of identifying purchase intent in UGC, with carrying out an analysis of the structure and content of posts and extracting feature from them. They make use of the linguistic preprocessors for feature extraction, such as dependency parser and named entity recogniser, which are only available for a handful of languages. [18] presented a graph-based learning approach to inferring intent categories for tweets. [18] primarily focus on identifying and classifying tweets that explicitly express user's purchase intentions. In order to prepare a training data with tweets that express user's purchase intentions, [18] proposed a bootstrap-based extraction model that made use of a seed list of purchase-indicators (e.g. 'want to'). They took help of manual annotators to classify the collected tweets into six different intent categories. The major disadvantage of these methods [7, 18] lies with their data set since their training data is based on a handful of keywords. In our work, we encountered this problem with employing an query expansion technique [14], which has essentially helped to increase the coverage of keywords in our training data. We are the first to train our models with deep learning technique (RNN with LSTM hidden units) for this problem, i.e. purchase intent identification task in social media platform.

3 Existing Dataset

This section details an existing labeled training data in which each tweet is associated with an appropriate purchase intent or non-intent category. The creation of this data set was based on a limited set of keywords. A brief overview of this dataset is provided below.

As mentioned in Section 2, [18] applied a bootstrapping based method to retrieve intent tweets from Twitter, given a seed set of intent-indicators, (e.g. 'want to'). A manual annotation process was carried out on those extracted tweets that contain at least one intent-indicator. In short, tweets were distinguished as intent and non-intent tweets and the intent tweets were categorised into six different categories, namely 'Food and Drink', 'Travel', 'Education and Career', 'Goods and Services', 'Event and Activities' and 'Trifle'. [18] shared their training data with us. From now, we call this data set *Dataset1*. The statistics of Dataset1 can be found in [18], which we also report in Table 1. As can be seen from Table 1, Dataset1 consists of 2,263 labeled tweets, with six intent and non-intent categories.

4 Our Dataset

This section details creation of a new training data. First, we explain why the existing data (i.e. Dataset1), to a certain extent, is inadequate for this task. Then, we demonstrate

	Total	Food & Drink	Travel	Career & Education	Goods & Services	Event & Activities	Trifle	Non-intent
Dataset1	2,263	245 11.50%	187 8.78%	159 7.46%	251 11.78%	321 15.07%	436 20.47%	531 24.92%

Table 1. The statistics of the existing training data set.

how we created a new dataset. The created dataset, in practice, is to be an ideal data set for addressing this problem.

4.1 Variation of intent-indicators

The expression of interest of a Twitter user may be associated with the user's state of mind, emotion, or other phenomenon. Hence, the different Twitter users can express their thoughts of interest in numerous ways. For example, the users may show theirs interest to purchase a product with any of the following intent-indicators: 'want to','need to','like to','wish to', and 'hope to'. Spelling mistake is a common phenomenon in tweets (e.g. short form, noisy long form). Hence, tweets may include misspelled intent-indicators. For example, when a user is really excited to buy a product soon, his purchase intent tweet can be 'I wannnntttt to buy an iPhone!!!!!!'. Similarly, intent indicators can be specified as 'n33d to', 'h0pe to', 'wnt to' and so on. All the prior studies in this direction do not take into the consideration of different ways by which an intent indicator can be specified. In this work, we aim to make the list of purchase intent indicators as exhaustive as possible, with taking the nature of the user generated contents in this media into consideration. In the next section we describe how we expand the existing list of seed purchase intend indicators.

4.2 Expanding the list of intent-indicators

As discussed above, the existing data set (i.e. Dataset1) has limited coverage of the indent-indicators. In order to increase the coverage, and to capture new tweets that are good paradigms of purchase intentions, we expand the list of intent-indicators[6] using a query expansion technique. This is accomplished with a continuous distributed vector representation of words using the continuous Skip-gram model (also known as Word2Vec) proposed by [14], by maximizing the objective function:

$$\frac{1}{|V|}\sum_{n=1}^{|v|}\sum_{-c\leq j\leq c, j\neq 0} logp(w_{n+1}|w_n) \tag{1}$$

where $|V|$ is the size of the vocabulary in the training set and c is the size of context window. The probability $p(w_{n+j}|w_n)$ is approximated using the hierarchical softmax.

[6] We get the intent-indicator list from [18]

Top-10 Similar Words/Phrases	Intent-indicator			
	want	need	wish	like
	wamt	neeed	Wish	lile
	wan't	nees	wished	likr
	wanr	neeeed	wishh	llike
	want/need	neeeeed	whish	likw
	wabt	meed	Wishhhh	lik
	wNt	Need	Wished	lkke
	wnat	n99ed	Wishin	Iike
	/want/	neex	WISH	lije
	need/want	need/want	Iwish	lke
	eant	neeeeeed	wishhh	lyk

Table 2. Top 10 similar words/phrases of the seed intend indicators: 'want', 'need', 'wish' and 'like'.

In our case, we used a pre-trained 200-dimensional GolVe vectors [15]. Given the vector representations for the intend-indicators, we calculate the similarity scores between words/phrase pairs in the vocabulary using cosine similarity. Top 10 similar words/phrase for each seed intent-indicator in our list are selected for the expansion of initial seed list. For an example, in Table 2, we list the top 10 similar intent-indicators of four seed intent-indicator: 'want', 'need', 'wish' and 'like'. Finally, in order to weed out irrelevant intent-indicators, if any, from the list of the expanded intent-indicators, a manual inspection was carried out.

4.3 Collecting tweets with the expanded seed intent-indicators

We extract tweets using a Pyhton library, Tweepy,[7] from Twitter. For extraction, we use the expanded list of seed intent-indicators (cf. Section 4.2). Potentially, each of the extracted tweets contains at least one intent-indicator. In Table 3, we show a few of those tweets that were collected from Twitter given the seed intent-indicator: 'n33d'.

I n33d yall to be more active on younow plz and thanks
I n33d more youtubers to watch
I n33d to make some fucking music
I n33d to inhale these pizz@ pringle
I am so hungry people got to watch out n33d foOOoOod
I n33d to stop buying jackets
I was at ritz last friday shown out y3w n33d to go out to da club wit m3
I think I need to go get some therapy

Table 3. A few of the tweets collected with the seed intent-indicator: 'n33d'.

[7] http://docs.tweepy.org/en/v3.5.0/api.html

4.4 Labeling collected purchase intent tweets

We randomly sampled a set of 2,500 tweets from the list of the collected tweets that contain at least one intent-indicator, and another set of 2,500 tweets from Twitter, each of them contains no intent-indicators. During sampling we ensure that none of the tweets overlaps with those from Dataset1 (cf. Section 3). Then, we applied a noise cleaning method on the tweets, i.e. all the null values, special characters, hashtags were removed from tweets. This cleaning process was carried out with a manual editor who has excellent English skills and good knowledge on tweets. After manual cleaning, we get a set of 4,732 tweets. As mentioned earlier in the paper, [18] defined a set of purchase intent categories (six) in order to classify those tweets that express users' purchase intention. Following [18] we label each of the collected clean tweets with tagging that with either one of purchase intent categories or non-intent category. The manual annotation process is accomplished with a GUI that randomly displays a tweet from the set of 4,732 tweets. The GUI lists the six intent (i.e. 'Food and Drink', 'Travel', 'Education and Career', 'Goods and Services', 'Event and Activities' and 'Trifle') categories and the sole non-intent category as in [18]. For the annotation purposes we hired three annotators who are native English speakers and have excellent knoweldge on UGCs (i.e. tweets). The annotators are instructed to follow the following rules for labeling a tweet:

- label each tweet with an appropriate category listed on GUI,
- skip a tweet for annotation if you are unsure about user's purchase intention in the tweet or its intention category,
- skip those tweets for annotation that are unclear and includes characters of other languages or noisy characters,
- label those tweets with the non-intent category that express negative intents (e.g., 'don't want to').

	Total	Food & Drink	Travel	Career & Education	Goods & Services	Event & Activities	Trifle	Non-intent
Dataset2	3,575	285 8.0%	214 6.0%	164 4.6%	387 10.8%	344 9.6%	450 12.6%	1,803 50.4%

Table 4. The statistics of the new training data set.

On completion of the annotation task, we obtained 4,210 tweets, each of which is associated with at least one tag.[8] Since we have three annotators and three values are associated with the most of tweets of the set of 4,210 labeled tweets, final class for a tweet is determined with two out of three voting logic. Thus, 635 annotated tweets were not considered in the final annotated set due to the disagreements of all three annotators. The final set of annotated tweets contains 3,575 entries. From now, we call this data set *Dataset2*, whose statistics are reported in Table 4.

[8] At least, one out of three manual annotators label each of the 4,210 tweets.

On completion of the annotation process, inter-annotator agreement was computed using Cohen's kappa [4] at tweet-level. For each tweet we count an agreement whenever two out three annotators agree with the annotation result. We found the kappa coefficient to be very high (i.e. 0.64) for the annotation task. This indicates that our tweet labeling task is to be excellent in quality.

4.5 Combined Training data

For our experiments we merged the training examples from Dataset1 with that of Dataset2. From now, we call the combined training set *ComDataset*. The statistics of ComDataset are reported in Table 5. For our experiments we randomly selected 1,000 examples from ComDataset, and create a test set with 500 examples and a validation set with 500 examples. The set of remaining 4,838 examples from ComDataset was considered as the training set.

	Total	Food & Drink	Travel	Career & Education	Goods & Services	Event & Activities	Trifle	Non-intent
ComDataset	5,838	530 9.1%	401 6.9%	323 5.5%	538 9.2%	665 11.4%	886 15.2%	2,336 40.0%

Table 5. The statistics of the combined training data (ComDataset).

5 Methodology

5.1 LSTM Network

Nowadays, RNN, in particular with LSTM [8] hidden units, has been proved to be an effective model for many classification tasks in NLP, e.g. sentiment analysis [19], text classification [11, 21]. RNN is an extension of the feed-forward NN, which has the gradient vanishing or exploding problems. LSTM deals with the exploding and vanishing gradient problems of RNN. An RNN composed of LSTM hidden units is often called an LSTM network. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. More formally, each cell in LSTM can be computed as follows:

$$X = \begin{bmatrix} h_{t-1} \\ x_t \end{bmatrix} \tag{2}$$

$$f_t = \sigma(W_f \cdot X + b_f) \tag{3}$$

$$i_t = \sigma(W_i \cdot X + b_i) \tag{4}$$

$$o_t = \sigma(W_o \cdot X + b_o) \tag{5}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh\,(W_c \cdot X + b_c) \tag{6}$$

$$h_t = o_t \odot tanh\ (c_t) \tag{7}$$

where $W_i, W_f, W_o \in \mathbb{R}^{d \times 2d}$ are the weighted matrices and $b_i, b_f, b_o \in \mathbb{R}^d$ are biases of LSTM, which need to be learned during training, parameterising the transformations of the input, forget and output gates, respectively. σ is the sigmoid function, and $\odot$ stands for element-wise multiplication. x_t includes the inputs of LSTM cell unit. The vector of hidden layer is h_t. The final hidden vector h_N represents the whole input tweet, which is passed to *softmax* layer after linearizing it into a vector whose length is equal to the number of class labels. In our work, the set of class labels includes intent and non-intent categories.

5.2 Classical Supervised Classification Models

Furthermore, we compare the deep learning model with the classical classification models. We employ the following classical supervised classification techniques:

- Baseline 1: Logistic Regression (LR)
- Baseline 2: Decision Tree (DT)
- Baseline 3: Random Forest (RF)
- Baseline 4: Naïve Bayes (NB)

These classical learning models (LR, DT, RF and NB) can be viewed as the baselines in this task. Thus, we obtain a comparative overview on the performances of different supervised classification models including LSTM network. Note that we consider the default set-up of an well-known machine learning library for our baseline classifiers (cf. Section 6).

6 Experiments

This section details the building of different classification models. In order to build LR, DT, RF and NB classification models, we use the well-known Scikit-learn machine learning library,[9] and performed all the experiments with default parameters set by scikit-learn. As for the representation space, each tweet was represented as a vector of word unigrams weighted by their frequency in the tweet. For building our neural network (NN) and training the model we use Lasagne library.[10] Our RNN model includes LSTM units. The size of input layer of the NN is 12,000. We employ layer normalisation [1] in the model. Dropout [5] between layers is set to 0.10. The size of embedding and hidden layers are 512 and 1024. The models are trained with Adam optimizer [12], with learning-rate set to 0.0003 and reshuffling the training corpora for each epoch. We use the learning rate warm-up strategy for Adam. The validation on development set is performed using cross-entropy cost function. The RNN model is trained up-to 20 epochs, and we set mini-batches of size 32 for update.

We observe the learning curve of the classification models with the following experimental set-up:

[9] https://scikit-learn.org/stable/

[10] https://lasagne.readthedocs.io/en/latest/

– Set-up 1: classifying tweets into the two classes: intent and non-intent. In this case, all intent sub-classes are merged into a one single intent class (cf. Table 5).
– Set-up 2: classifying tweets into seven classes: six intent categories ('Food and Drink', 'Travel', 'Education and Career', 'Goods and Services', 'Event and Activities' and 'Trifle') and one non-intent category.
– Set-up 3 (one vs all): in this set-up we select a particular intent class, and the remaining intent sub-classes are merged into one single class. Like the first set-up (Set-up 1), this one is a binary classification task. In order to test classifiers in this set-up, we chose the following two intent classes: 'Goods and Services' and 'Trifle'.

6.1 Results and Discussion

	Precision	Recall	F_1-score	Precision	Recall	F_1-score
	LR			DT		
Intent	0.79	0.89	0.82	0.75	0.81	0.78
Non-intent	0.88	0.73	0.80	0.80	0.74	0.77
avg/total	0.82	0.81	0.81	0.78	0.77	0.77
	RF			NB		
Intent	0.76	0.89	0.82	0.76	0.89	0.82
Non-intent	0.87	0.73	0.79	0.88	0.73	0.80
avg/total	0.82	0.81	0.81	0.82	0.81	0.81
	RNN					
Intent	0.82	0.86	0.84			
Non-intent	0.88	0.77	0.83			
avg/total	0.85	0.82	**0.83**			

Table 6. Accuracy of classification models (set-up 1: intent and non-intent) measured with precision, recall and F_1-score metrics.

We evaluate the performance our classifiers and report the evaluation results in this section. In order to measure classifier's accuracy on the test set, we use three widely-used evaluation metrics: precision, recall and F_1-score. Note that we could not directly compare the approach of [18] with ours since the source code of their model is not freely available to use. We report the evaluation results obtained with the experimental set-up 'Set-up 1' (cf. Section 6) in Table 6. Here, we draw a number of observations from the evaluation results presented in Table 6:

1. We see excellent performance with all classifiers for both intent and non-intent catergories.
2. Irrespective of the classification models, the accuracy of identifying purchase intent tweets is slightly better than that of identifying non-intent tweets.
3. When we compare the scores of different classification models on the test set, we see that the RNN model becomes the winner, with achieving a F_1-score of 0.83 on the gold-standard test set.

	Precision	Recall	F_1-score	Precision	Recall	F_1-score
	LR			DT		
Food and Drink	0.87	0.83	0.85	0.54	0.42	0.30
Travel	0.69	0.57	0.62	0.34	0.37	0.36
Education and Career	1.0	0.51	0.68	0.54	0.54	0.54
Goods and Services	0.77	0.56	0.65	0.60	0.57	0.59
Event and Activities	0.71	0.51	0.68	0.29	0.31	0.30
Trifle	0.47	0.45	0.46	0.35	0.47	0.40
Non-intent	0.78	0.94	0.86	0.80	0.76	0.78
Avg/Total	0.75	0.75	0.74	0.63	0.61	0.62
	RF			NB		
Food and Drink	0.61	0.69	0.65	1.0	0.12	0.22
Travel	0.49	0.54	0.51	1.0	0.03	0.06
Education and Career	0.71	0.57	0.63	1.0	0.06	0.11
Goods and Services	0.62	0.71	0.66	0.89	0.18	0.30
Event and Activities	0.54	0.28	0.37	1.0	0.03	0.05
Trifle	0.35	0.34	0.34	0.43	0.04	0.07
Non-intent	0.80	0.84	0.82	0.54	1.00	0.70
Avg/Total	0.67	0.68	0.67	0.69	0.55	0.43
	RNN					
Food and Drink	0.90	0.85	0.88			
Travel	0.76	0.68	0.72			
Education and Career	0.74	0.73	0.74			
Goods and Services	0.86	0.67	0.75			
Event and Activities	0.92	0.96	0.94			
Trifle	0.57	0.54	0.55			
Non-intent	0.67	0.91	0.77			
Avg/Total	0.77	0.76	**0.76**			

Table 7. Accuracy of classification models (set-up 2) on six intent classes and one non-intent class measured with precision, recall and F_1-score metrics.

Next, we report the evaluation results obtained with the experimental set-up 'Set-up 2' (cf. Section 6) in Table 7. This time, the classification task involves seven output classes, i.e. six intent classes and the sole non-intent class. Here, we draw a number of observations from the evaluation results presented in Table 7:

1. In general, we get high precision and low recall scores for the intent categories. For the non-intent class, most of the cases, the scenario is the other way round.
2. As far as the scores with the F-score metric are concerned, the performances of RNN and LR classifiers look reasonable, and the performances of the remaining classifiers (i.e. DT, NB and RF) look moderate.
3. When we compare different classification models, we see, as in Set-up 1, the RNN model becomes the winner, with achieving F_1-score of 0.76 (average) on the gold-standard test set.
4. As can be seen from Table 7, the recall scores of NB classifier are below par, and even in some cases, those are very poor. We recall Table 5 where we can see the

presence of class imbalance in the training data. For instance, 6.9% and 5.5% training examples belong to 'Travel' and 'Career and Education' classes, respectively. This could be one of the reasons why classifiers performed poorly for some categories. This phenomenon is also corroborated by Gupta et al. [7] who built classifiers with a training data having the class imbalance issues.

	Precision	Recall	F_1-score	Precision	Recall	F_1-score
	LR			DT		
Goods and Services	0.88	0.76	0.82	0.80	0.76	0.78
Intent	0.90	0.96	0.93	0.90	0.92	0.91
avg/total	0.90	0.90	0.90	0.87	0.87	0.87
	RF			NB		
Goods and Services	0.80	0.72	0.76	0.88	0.47	0.61
Intent	0.89	0.92	0.91	0.81	0.97	0.89
avg/total	0.86	0.86	0.86	0.83	0.82	0.80
	RNN					
Goods and Services	0.92	0.98	0.95			
Intent	0.94	0.98	0.96			
avg/total	0.93	0.98	**0.95**			

Table 8. Accuracy of classification models (set-up 3: 'Goods and Services' and a combined class from the rest of the intent categories measured with precision, recall and F1 metrics.

Now, we observe the learning curve of the classifiers with the third experimental set-up (i.e. 'Set-up 3', cf. Section 6) where a particular intent category (e.g. 'Goods and Services'/'Trifle') is held and the rest of the intent categories are merged into a single category. In this set-up, we remove those examples from the test and development sets that include the non-intent target class. We test our classifiers on the test set and obtain the evaluation results. We report the evaluation results for 'Goods and Services' in Table 8 and for 'Trifle' in Table 9. Here, we draw a number of observations from the evaluation results presented in Table 8 and 9:

1. We see an excellent performance across the classifiers for 'Goods and Services' and the combined intent category.
2. We get a mix bag of results across the classifiers and metrics for 'Trifle' and the combined intent category .
3. Like the above experimental set-ups, in this set-up, the RNN models become superior than the other classification models in identifying users' purchase intent type in tweets. The RNN model produces an accuracy of F_1-score of 0.95 (average) on the test set when 'Goods and Services' category is considered. As far as the 'Trifle' category is concerned, the RNN model gives an accuracy of F_1-score (average) of 0.83 on the test set.
4. When we consider the recall scores in Table 9, we see most of the classifiers performed below par with the 'Trifle' category.

	Precision	Recall	F_1-score	Precision	Recall	F_1-score
	LR			DT		
Trifle	0.70	0.43	0.54	0.53	0.45	0.49
Intent	0.83	0.94	0.88	0.82	0.86	0.84
avg/total	0.80	0.81	0.79	0.75	0.76	0.75
	RF			NB		
Trifle Trifle	0.66	0.38	0.48	1.00	0.09	0.17
Intent	0.81	0.93	0.87	0.76	1.00	0.87
avg/total	0.77	0.79	0.77	0.82	0.77	0.69
	RNN					
Trifle	0.88	0.69	0.77			
Intent	0.93	0.87	0.89			
avg/total	0.91	0.78	**0.83**			

Table 9. Accuracy of classification models (Set-up 3: 'Trifle' and a combined class from the rest of the intent categories measured with precision, recall and F1 metrics.

7 Conclusion

In this paper, we presented supervised learning models to identify users' purchase intent from the tweet data. We present the first study to apply LSTM network for purchase intent identification task. With our RNN classifiers we achieved reasonable accuracy (F_1-scores ranging from 0.76 to 0.95) in all classification tasks. This shows the applicability of deep learning algorithms to a classification task where a tiny training data is available. Further, we demonstrated the efficacy of the LSTM network by comparing its performance with different classifiers. The major portion of the paper describes the way we created our own training data. The existing training data set for this task was not satisfactory as it is limited with a set of keywords. We semi-automatically created training data set, with employing a state-of-the-art query expansion technique [14]. This has essentially helped to increase the coverage of keywords in our training data.

In future, we intend to make our gold standard data set available to the NLP community. We also plan to test our method on different social media platform, e.g. Facebook,[11] and with different languages. We also intent to apply our methods to a cross-lingual social platform. We plan to increase the size of training examples for those classes for which we have lesser proportion of training examples. This could encounter the class imbalance problem in our training data. All the resources developed in this current study including dataset and extended list of intent indicators will be released publicly on publication and can be downloaded from `www.anonymous.com`.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. CoRR **abs/1607.06450** (2016), `https://arxiv.org/abs/1607.06450`

[11] https://www.facebook.com/

2. Beitzel, S.M., Jensen, E.C., Lewis, D.D., Chowdhury, A., Frieder, O.: Automatic classification of web queries using very large unlabeled query logs. ACM Trans. Inf. Syst. **25**(2) (Apr 2007). https://doi.org/10.1145/1229179.1229183, http://doi.acm.org/10.1145/1229179.1229183
3. Cao, H., Hu, D.H., Shen, D., Jiang, D., Sun, J.T., Chen, E., Yang, Q.: Context-aware query classification. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. pp. 3–10. ACM (2009)
4. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20**(1), 37–46 (1960)
5. Gal, Y., Ghahramani, Z.: A theoretically grounded application of dropout in recurrent neural networks. CoRR **abs/1512.05287** (2016), https://arxiv.org/abs/1512.05287
6. Goldberg, A.B., Fillmore, N., Andrzejewski, D., Xu, Z., Gibson, B., Zhu, X.: May all your wishes come true: A study of wishes and how to recognize them. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 263–271. Association for Computational Linguistics (2009)
7. Gupta, V., Varshney, D., Jhamtani, H., Kedia, D., Karwa, S.: Identifying purchase intent from social posts. In: ICWSM (2014)
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
9. Hu, J., Wang, G., Lochovsky, F., Sun, J.t., Chen, Z.: Understanding user's query intent with wikipedia. In: Proceedings of the 18th international conference on World wide web. pp. 471–480. ACM (2009)
10. Jansen, B.J., Booth, D.L., Spink, A.: Determining the informational, navigational, and transactional intent of web queries. Information Processing & Management **44**(3), 1251–1266 (2008)
11. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)
12. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014), http://arxiv.org/abs/1412.6980
13. Li, X., Wang, Y.Y., Acero, A.: Learning query intent from regularized click graphs. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 339–346. ACM (2008)
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
15. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
16. Ramanand, J., Bhavsar, K., Pedanekar, N.: Wishful thinking: finding suggestions and'buy'wishes from product reviews. In: Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text. pp. 54–61. Association for Computational Linguistics (2010)
17. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. nature **323**(6088), 533 (1986)
18. Wang, J., Cong, G., Zhao, W.X., Li, X.: Mining user intents in twitter: A semi-supervised approach to inferring intent categories for tweets. In: AAAI. pp. 318–324 (2015)
19. Wang, Y., Huang, M., Zhao, L., et al.: Attention-based lstm for aspect-level sentiment classification. In: Proceedings of the 2016 conference on empirical methods in natural language processing. pp. 606–615 (2016)

20. Werbos, P.J.: Backpropagation through time: what it does and how to do it. Proceedings of the IEEE **78**(10), 1550–1560 (1990)
21. Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., Xu, B.: Text classification improved by integrating bidirectional lstm with two-dimensional max pooling. arXiv preprint arXiv:1611.06639 (2016)