ISSN 2278-3091
Volume 9, No.3, May - June 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse214932020.pdf
https://doi.org/10.30534/ijatcse/2020/214932020

Split Over-Training for Unsupervised Purchase Intention Identification

Noor Fazilla Abd Yusof[1], Chenghua Lin[2], Xiwu Han[3], M Hardyman Barawi[4]
[1]Centre for Advanced Computing Technology (C-ACT), FakultiTeknologi Maklumat dan Komunikasi (FTMK),UniversitiTeknikal Malaysia Melaka, Malaysia, elle@utem.edu.my
[2]Computer Science Department, University of Sheffield, United Kingdom, c.lin@sheffield.ac.uk
[3]School Of Arts, English and Languages, Queen's University Belfast, United Kingdom, x.han@qub.ac.uk
[4]Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak, Malaysia, bmhardyman@unimas.my

ABSTRACT

Recognizing user-expressed intentions in social media can be useful for many applications such as business intelligence, as intentions are intimately linked to potential actions or behaviors. This paper focuses on a binary classification problem: whether a text expresses purchase intention (PI) or not (non- PI). In contrast to existing research, which relies on labeled intention corpus or linguistic knowledge, we proposed an unsupervised method called split over-training for the PI identification task. Experiments on PI identification from tweets showed that our approach was effective and promising. The best classifying accuracy of 84.6% and PI F-measure of 70.4% was achieved, which are only 7.7% and 4.9% respectively lower than fully supervised models. This means our unsupervised method may provide reasonable preprocessing for intention corpus labeling or intention knowledge acquisition.

Key words : Intention analysis, text analysis, purchase intention identification

1. INTRODUCTION

Social network analysis has gained significant attention in recent years, largely due to the success of online social networking platforms, and the consequent availability of a wealth of social network data. One interesting and important aspect of making use of and understanding such valuable data is to identify the intention expressed in user-generated messages[1]. Formally, an intention or intent, which often involves planning and forethought, is a mental state that represents a commitment to carry out an action or actions in the future[2]. Intentions are intimately linked to potential actions or behaviors e.g. continuance intention to use Facebook[3], intention of using mobile apps in transportation (i.e. GoJek and Grab)[4] etc. The ability of recognizing and understanding intentions from texts is also crucial for many domains like business intelligence[1], cyber security[5], and industrial robotics[6]. For instance, the tweet "*Anyone have a suggestion in Paris for a 2-night getaway idea*? *Coast, mts*? *Need ideas*!" contains a purchase intention (PI), i.e., booking accommodation in Paris. Automatically detecting purchase intentions expressed in texts can be very useful for business intelligence as the intentions may directly correspond to an immediate business opportunities or admiration for products or services.

Existing research about intention identification are either based on acquired linguistic knowledge of syntactic and/or semantic patterns[7], [8] or manually labeled intention corpus[1], [5]. To our knowledge, there is no large-scaled labeled social media intention corpus or linguistic knowledge database publicly available. However, the booming volume (e.g. 500 million tweets per day[1] and wide variety of social media texts (e.g. posts in Facebook are often both linguistically and pragmatically different from Tweets) make the acquisition of linguistic knowledge greatly challenging, and large corpus labeling very expensive.

Motivated by the above observations, and under the linguistic hypothesis that PI and non-PI texts are formally distinguishable, we proposed a novel unsupervised algorithm called split over-training for identifying purchase intentions from social media. Our algorithm analytically borrowed the concept of split over-training from athletic science[9]. Both athletic split over-training and our algorithm explore and benefit from the relations among parts of the main body/corpus, and relations between the main body/corpus and each part. Our algorithm is distinguished from existing approaches for intention detection in three aspects: a) free from corpus labeling; b) independent from domains or social media sources for training; and c) complementary to supervised approaches.

[1]http://www.internetlivestats.com/twitter-statistics/

We experimented on purchase intention identification from Twitter texts represented on three levels: whole tweets, local patterns of word windows around general intention words, and local patterns of dependency structures around general intention words. Evaluations on 2K manually annotated tweets showed that our proposed method was effective and promising, with the best OCA (overall classification accuracy) of 83.8% and F-PI (F-measure score for the PI class) 68.2% achieved on 3-level local dependency structures. In addition, we linearly combined models trained on different corpus representations, and achieved even better performances, with the best OCA of 84.6% and F-PI of 70.4%, which are only 7.7% (OCA) and 4.9% (F-PI) lower than the upper bound based on a fully supervised model trained on the testing set.

Furthermore, the state-of-the-art supervised method[1] produced OCA of 92.8% and F-PI of 76.5% after being fitted on the testing set, which is comparable to our upper bound. This proves our hypothesis: PI and non-PI texts are indeed formally different in general, and even more distinguishable from each other locally around general intention words.

2. RELATED WORKS

In this section, we first address the relevant theories of intentions in section 2.1. We then discuss the existing research on understanding intentions in section 2.2.

2.1 Relevant Theories

In philosophy, due to the different renditions, intention has been approached as: a) action with which another action is done in terms of a primary reason[10]; b) action in terms of acting for a reason[11]; and c) practical attitude marked by its pivotal role in planning for the future[2]. Cognitive or robotic research often adopt one or both of the first two renditions, and focus on the process of becoming aware of the intention of another agent or inferring an agent's intention through its actions and their effects on the environment [6], [12]–[14]. Social media or web-based research usually take the last rendition, and try to detect intention as practical attitude from texts[1], [5], [15]. Our research complies with the last rendition.

In Speech Act Theory, performative utterances or sentences involve locutionary, illocutionary, and perlocutionary acts [16]. For an utterance, there are usually “primary” and “secondary” illocutionary acts simultaneously, with the primary also called an indirect speech act and the secondary a direct one[17]. Some research classify emails or sentences directly into different speech acts[7], [18], [19], [20],while others and our present work focus on the speech acts that communicate future intentions [1], [5], [15]. Specifically, we regard linguistically instantiated or textual intentions as indirect illocutionary speech acts with special reasons or practical attitudes.

Relevant communication theories depict intentions as sensitive to specific scenarios or contexts, across which the speaker's intention or the intention inferred by the receiver or an audience may vary widely[7]. The context sensitivity makes manual annotation of intention corpus more difficult and expensive, which is one of our motives for pursuing the unsupervised approach. In a commercial survey, questionnaires about purchase intention often employ scalable expressions to form alternative answers to intention questions, and these scales are then translated into intention probabilities[21]. In this study, we use the prediction confidence of classifiers to simulate the intention probability.

2.2 Computational Efforts

Existing research on intention identification or analysis from natural language involve different social media platforms, various textual genres, and a diversity of intention categories. Mostly covered social media includes Facebook[22], Yahoo[8], Twitter[15]and Amazon[5]; textual genres include emails[7], chat rooms[23] and message boards[18]; intention categories vary from two simple ones as commercial versus non-commercial[24] to very complicated 136 intention types from a social-psychological framework[8].

There have been generally three types of approaches for identifying intention: i) based on keyword ontology[7]or intention knowledge[8]; ii) linguistically data-driven expression features[18]or syntactic/semantic intention patterns[5]; and iii) supervised machine learning approaches using manually labeled intention corpus[1]. To the best of our knowledge, there is no large-scaled labeled social media intention corpus or linguistic knowledge database publicly available, and the booming volume and wide variety of social media texts can make the acquisition of linguistic knowledge greatly challenging, and large corpus labeling very expensive.

Therefore, this study proposes an unsupervised method for identifying purchase intentions from social media texts, which requires no other linguistic or domain knowledge for model training except a short list of manually collected intention keywords and the hypothesis that PI and non-PI texts are formally distinguishable. Furthermore, the unsupervised characteristic makes our method independent from social media sources and intention categories. [5]manually built a domain-independent framework of intention lexicon and grammar for general intention identification. The limited coverage and difficulty in extending are their disadvantages, which might be avoided with our unsupervised method.

2.3 Formulations and Definitions and Acronyms

In this section, we define the relevant concepts and problems for the task of purchase intention (PI) identification from social media by drawing on the experiences and conventions in sentiment analysis[25][26].

A. Intention

An intention in social media is defined as a quintuple $(a_i, C(W_{ijk}), e_j, t_k, p_{ijk})$,where a_i is an intention holder;
W_{ijk} is one or a group of intention words or phrases; $C(W_{ijk})$ maps W_{ijk} to a set of predefined intention categories C; e_j is a target entity or intention theme; t_k is the time stamp when the intention is expressed; and p_{ijk} is the degree of a_i s certainty about this intention or the probability of the social media text being considered as an intention. The tweet (a) in Table 1expresses the purchase intention, where can be formatted as $(a_i = 'x@y', C(W_{ijk}) = C(\{'will', 'buy'\}) = 'to\ buy', e_j = 'black\ coffe', t_k = '11/22/201421{:}38{:}03', p_{ijk} = 1.0)$.

Table 1: Examples of Purchase Intention in Twitter.

(a)	x@y (11/22/2014 21:38:03): I'll definitely buy someblack coffee for my best!!!
(b)	I want some black coffee for my best.
(c)	My best is thirsty.

B. General Intention Part (GI) and Special Intention Part (SI)

The intention words or phrases in W_{ijk} may include a General Intention Part (e.g. '*will*', '*want*', '*need*', etc.) that functions as a lexical symbol for the intention, and a Special Intention Part (e.g. '*buy*', '*purchase*', '*drink*', etc.) that specifies the intended immediate or future action. Both GI and SI can be optional, but SI is often collocated with GI. As shown in the tweet (a) in Table 1, GI = '*will*', and SI = '*buy*', where W_{ijk} is made up of both GI and SI.

C. Explicit Intention vs. Implicit Intention

An explicit intention is consistent with the literal meaning or locutionary/direct speech act described in the text. The intention example of (a) in Table 1 is explicit if we define C as {'*to buy*','*not to buy*'}. GI is indispensable for W_{ijk} in explicit intentions. An implicit intention lies in the hidden meaning or illocutionary/indirect speech act, of which the direct speech act can either be mapped to another intention or an informative text about a fact that may reasonably lead to the implicit intention. For the same intention, an implicit expression usually results in smaller p_{ijk} than an explicit expression, and p_{ijk} is also sensitive to the degree of implicitness. For the intention of'*buy coffee*', sentences of (b) and (c) in Table 1 might be considered of respectively less certainty in the speaker's mind, and thus are examples of being implicit.

D. Purchase Intention (PI)

A purchase intention is a text expression signifying a desire or need to purchase or consume a product or service [15], and is closely related to product recommendation[27], [28]. Specifically, SI in a PI is either a word with purchasing meaning (i.e. explicit) or indicates a purchasing potentiality (i.e. implicit). All examples in Table 1 could be regarded as PIs.

E. Intention Identification

The task of intention identification (aka. intention analysis or mining) can be understood as identifying a corresponding intent category for every action indicative in a given text [8]and/or specifying the five elements in the quintuple. The present work focuses on the first part of this task, which is to approximate the prediction function $f: S \times C^I \Rightarrow [0,1]$, where $S$is a text or a sequence of sentences, $C^I = \{C_1^I, C_2^I, \dots, C_n^I\}$is a set of predefined intent categories, and $[0,1]$defines the value range of p_{ijk}.

F. Purchase Intention (PI) Identification

Here, we address the task of identifying PI from social media texts, or the two-class classification problem where $C^I = \{PI, non - PI\}$. Specifically, we regard a text expression S as PI when $f(S \times C^{PI}) - f(S \times C^{non-PI}) = \delta > 0$, where $\delta$is the prediction margin of a classifier about S, and $\theta$is an empirical threshold.

3. METHODOLOGY

Instead of manually labeling corpus or acquiring linguistic knowledge, we propose split-over training, an unsupervised approach to the task of PI identification from social media texts. Flow chart in Figure 1 illustrates the whole process, which takes social media texts as input and outputs a final classifier to label new texts. There are two major parts involved in the process, i.e. clustering and classification. The clustering part implicitly exploits possible formal differences among the input social media texts, and the classification part iteratively develops the clustering results into a final classifier by means of split over-training.

First, a large corpus $D$with Nentries is retrieved from social media by filtering the raw input texts with a given list of GI words; second, the N entries are then clustered respectively into 2 groups and Kgroups, where K should be comparatively much larger than m, which is the number of splitting on the dataset, and is in positive proportion to N; for the 2-cluster results, the group with more entries that contain words from a given list of SI are labeled as possible PIs, and the other group as possible non-PIs; finally, a baseline classifier f' is trained on the labeled 2-cluster results. $f'$and $D$with K-clustering results will be input to the classification part of split over-training.

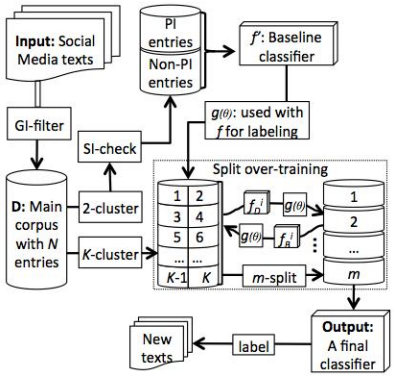

Figure 1: Flow Chart for Unsupervised PI Identification from Social Media Texts.

Table 2: Pseudocode for Split Over-Training.

Input: D the main corpus with k-clustering results; f' a baseline classifier; θ an empirical threshold.
Output: A final classifier for labeling new texts.

Algorithm:
Step 1: Randomly split D into m blocks such that each block B_j contains N/m entries;
Step 2: Predict the labels for all the N entries in D with f' and $g(\theta)$;
Step 3: Train a new classifier f_D^1 on D with the newly predicted labels;
For each block B_j of the m blocks, set $i = 1$
Loop:
Step 4: Predict labels for all entries in B_j with $f_D{}^i$ and $g(\theta)$;
Step 5: Train a classifier $f_B{}^i{}_j$ on B_j with newly labeled entries;
Step 6: Go to Step 10 if the cross-validation performance $F\left(f_B{}^i{}_j\right)$ deteriorates, or $f_B{}^i{}_j$ overfits, or i goes beyond a given limit $MaxI$;
Step 7: Set $\theta = \theta - \Delta i$;
Step 8: Predict the labels for all entries in D with $f_B{}^i{}_j$ and $g(\theta)$;
Step 9: Set $i = i + 1$, train a new classifier f_D^i on D with newly labeled entries, and go to Step 4;
End loop;
Step 10: Predict labels for entries in B_j with $f_B{}^{i-1}{}_j$;
End for;
Step 11: Train the final classifier on $B\{1, 2, \dots, m\}$.

Function $g(\theta)$: (Refer also to Figure 2)
Step 1: For a given classifier f and each entry S, if when $f(S \times C^{PI})$- $f(S \times C^{non-PI}) > \theta$, label S as PI; else label S as non-PI;
Step 2: For each cluster k of the K clusters, if in k the proportion of entries labeled as PI in Step 1 is larger than θ, re-label all entries in k as PI; else if in k the proportion of entries labeled as non-PI in Step 1 is larger than θ, re-label all entries in k as non-PI.

In athletic science, split over-training is a kind of planned over-training[9], which splits the training program so that different sets of muscles are worked on different days. We analytically borrowed this concept to describe our algorithm that over-trains a PI/non-PI classifier by iteratively exploiting the formal differences and relations among social media texts. Table 2 gives the pseudocode for the split over-training algorithm. The initial value of θ is related to PI and non-PI distribution of the corpus, but this relation is not very sensitive. Since θ is set for the PI proportion, which tends to be much smaller than that of non-PI in social media, it could be set very close to 1 initially and decreased by Δi, where $lim_{i\to\infty} \Delta = 0$. Function $g(\theta)$ is the key module for exploiting the clustering results by fine-tuning the predicted labels of a classifier f. The tuning process of $g(\theta)$ is graphically explained in Figure 1.

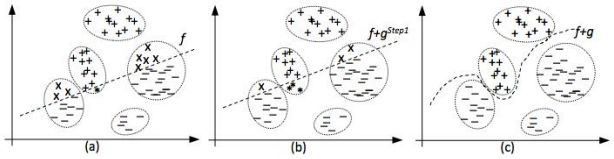

Figure 2: The Function of $g(\Theta)$: (a) Labeling Before Step 1; (b) Labeling after Step 1; (c) Labeling after Step 2 (Notes: +: True Positive; -: True Negative; x: False Positive; *: False Negative).

The over-training stops in Step 6 when the cross-validation performance $F\left(f_B{}^i{}_j\right)$ deteriorates, or $f_B{}^i{}_j$ overfits, i.e. $1 - F\left(f_B{}^i{}_j\right) < \varepsilon (NB. \varepsilon > 0)$, or i goes beyond a given limit $MaxI$. The first two halting conditions are closely related to practical purposes. For PI identification, a comparatively large ε and F-measure score for the PI class (F-PI) could be a good $F\left(f_B{}^i{}_j\right)$, while a smaller ε and the percentage of correctly classified entries (OCA) serves better preprocessing corpus for manual annotation or knowledge acquisition. $MaxI$ is negatively related to Δ. A larger Δ usually leads to smaller $MaxI$, and vice versa.

4. EXPERIMENTAL SETUP

As Twitter is one of the most popular social media platforms with 284 million monthly active users and 500 million tweets sent per day, we test our algorithm for the PI identification task on a corpus of collected tweets.

4.1 Dataset

Our intention lexicon involves a manually collected short list of 12 general intention key-words GI = {'*want*', '*need*', '*will*', '*plan*', '*intend*', '*eager*', '*purpose*', '*goal*', '*aim*', '*interest*', '*look for*', '*look forward to*' }, which conveys some degrees of intentionality, and 4 special intention keywords SI = {'*purchase*', '*buy*', '*pay*', '*order*'}, which describes the action to purchase and were used to determine the class labels of the 2-clustering results. We collected as raw input data about 1.4M tweets using the Twitter streaming API[2]. The raw data were then filtered with GI, i.e. a tweet was retrieved if it contained any of the 12 general intention keywords, as the same way by[5]. The final dataset consisting of 142,50 tweets, from which 2,000 were randomly held out to serve as the test set, where the average length of a tweet is 19.1 words. The remaining 12,250 tweets were used for split over-training, and the average length is 17.5 words.

The language used in social media, especially in tweets, is often informal and lacks overall grammatical structure; usage of acronyms and miss-spellings is also common. These factors make the social media data highly dimensional in

[2] https://dev.twitter.com/streaming/userstreams

nature, and the semantic analysis of social media difficult [15].At the sub-sentence level, however, words in social media texts tend to be arranged in correct order and grammatically to a certain degree[29]. Therefore, it is also important for our task to answer the research question: How PI and non-PI texts are formally different from each other and related among themselves. The answer to this question depends on how effectively our unsupervised approach will perform.

Besides using the whole tweets as main corpus, we also experimented separately with the local patterns of n-word window and n-level dependency around GI keywords as input data. A local pattern of n-word window is defined as a sequence $w_n = (w_{-j}, \dots, w_{-2}, w_{-1}, w_0, w_1, w_2, \dots, w_k)$, where w_0 is the GI keyword, and $n = j + k + 1$. The English tweet parser TWEEBOPARSER [3] [30] was employed for obtaining the dependency structures of tweets. A local pattern of n-level dependency is defined as a sequence $d_n = (s_{-j}, \dots, s_{-2}, s_{-1}, s_{-0}, s_1, s_2, \dots, s_k)$, where s is a sub-sequence of words on the same dependency level, s_0 is the sub-sequence of the GI keyword, $s\{_{-j}, \dots, _{-2}, _{-1}\}$ are sentential constituents indirectly or directly governing s_0, $s\{_{1,2}, \dots, _k\}$ are sentential constituents indirectly or directly governed by s_0, and $n = j + k$. For the tweet and its dependency structure in Figure 3, where '*want*' is GI, some examples are w_5 = "*I would want some black*", w_6 = "*I would want some black coffee*",d_2= "*I would want coffee for*", and d_3 = "*I would want some black coffee for my best*". Because there can be more than one GI keywords in one tweet, one or more local patterns may be extracted. To remove possible redundancy, the same GI won't appear in two or more patterns in adjacency to another GI.

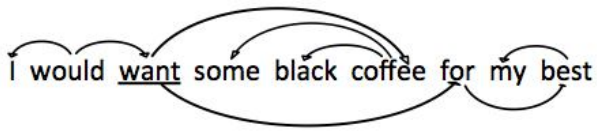

Figure 3: An Example of Dependency Structure.

4.2 Settings

We prepared five different versions of main corpus D, as listed in Table 3. Since local patterns of other lengths lead to rapid overfitting of ${f_B}^i{}_j$, we do not report them in this study. We employed the clustering and classification APIs in Weka 3.7[4] for all experiments. Simple K-Means was used for 2 and$K$clustering. TF and IDF parameters were set to be 'true', while other parameters were kept default. Naive Bayes Multinomial for Text (NBMT) was used for training all classifiers, and 10-fold cross validations were performed to provide halting measurements for over-training.

Table 3: Different Versions of the Main Corpus

Main corpus D	Number of entries
w_t: whole tweets	12,250
w_5: 5-word window	18,860
w_6: 6-word window	18,860
d_2: 2-level dependency	18,580
d_3: 3-level dependency	18,580

For the algorithm of split over-training, there are five parameters to be set, which are listed in Table 4. Our experiments showed that on all the five different versions of main corpus $m$with a value between 5-10 made no significant difference that could affect the convergence in the training process, while other smaller or larger numbers might lead to rapid overfitting of${f_B}^i{}_j$. We report results of 5- and 10-fold split over-training.

Table 4: Parameter settings of Split Over-Training.

Parameters	Settings
m for random split	5 and 10
Initial value of θ for $g(\theta)$	0.98
Descending value ofΔ for $g(\theta)$)	1.6/(i+3)(i+4)
$F\left({f_B}^i{}_j\right)$for halting conditions	F-PI or OCA
$\epsilon$for overfitting measurement	0.05

We set $\theta = 0.98$, which implies$f(S \times C^{PI}) > 0.99$. There are two reasons for such a setting. First, according to our observation, the classification model of Naive Bayes Multinomial for Text tends to assign large confidence to all its prediction; second, in our main corpus PI entries are much fewer than non-PI entries, and the estimated ratio of PI : non-PI could be less than 1:5. In $\Delta = 1.6/(i + 3)(i + 4)$, $i$is the iteration index of over-training. This setting made the convergence quick (often within 5 or 6 iterations for one split) yet without rapid overfitting. Another advantage of this setting is that it ensures $\theta > 0.5$ when $i \leq 10$, which is suitable for the small proportion of PI entries in the main corpus. We experimented with both F-PI and OCA as halting measurements for the purposes of practical PI identification and corpus annotation in our future work. $\epsilon = 0.05$means ${f_B}^i{}_j$is considered as over-fit when$F\left({f_B}^i{}_j\right) \geq 0.95$. Besides, there are also two relevant peripheral parameters: $K$and$MaxI$, which do not influence the training process directly. In fact, it was found that the setting $K \geq 1.5m$works for the algorithm very well. We set$K$to be 20 simply because for our data size it's the largest number to which Weka API provides easy memory management.$MaxI$is relevant toΔ, the above setting of which ensured$MaxI \leq 10$.

[3] http://www.ark.cs.cmu.edu/TweetNLP/#parser_down
[4] http://www.cs.waikato.ac.nz/ml/weka/

5. EXPERIMENTAL RESULTS AND EVALUATION

We implemented two groups of experiments respectively with OCA (overall classifying accuracy) and F-PI (F-measure score for the PI class) as $F\left(f_B{}^i{}_j\right)$, i.e. the convergence or overfitting measurement for over-training on each random split B_j. For both groups, NBMT classifiers were developed on the five different versions of the main corpus listed inTable 3 with 5 and 10 random split over-training. NBMT classifiers trained on 2-cluster and SI-check results served as baseline[5], and the NBMT classifiers trained on respective forms of the manually labeled testing set served as upper bounds for comparison. Besides, we also combined the three classifiers of 5 split over-training on w_t, w_6 and d_3 to further explore their potentialities (Refer to Figure 4).

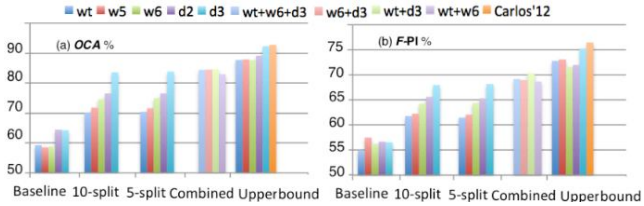

Figure 4: Evaluation Results on the Testing Set.

In order to compare our methods with the state-of-the-art research, we also fitted on our test set a rule-based classifier: Carlos'12, on the basis of the intention analysis results by [1]. Their method involves supervised Naive Bayesian, Maximum Entropy and SVM classification models, which able to detect multi-intention, sentiment and negation. It detected 15 intentions (See Table 5) in our testing set, and assigned them with sentiment values of '*positive*', '*neutral*' and '*negative*', and negation values of '*negated*' and '*not-negated*'.

Table 5: Intention Categories [1].

accuse	suggest	purchase	other	sell
direct	inquire	complain	thank	meet
opine	compare	apologize	quit	wish

As our task is to detect purchase intention, we manually developed 9 rules to map the 15 intention categories to PI and non-PI. Thus, Carlos'12 can detect implicit as well as explicit purchase intentions. These mapping rules were built in a way to fit an optimal classifier on our test set, so the ability of Carlos'12 to generalize may not be good but this disadvantage doesn't affect its function as one more upper bound comparison for our unsupervised method. Table 6 lists the 8 rules that map intentions of[1] to PIs, and the other rule not covered here maps all the other conditions to non-PIs.

[5] Another naive and straightforward baseline performance might be the SI-checking results on the testing set, for which OCA = 71.6% and F-PI = 8.2%.

Table 6: Purchase Intention (PI) Mapping Rules.

Wish→ PI	accuse (negated) →PI
compare → PI	apologize (negated) → PI
inquire →PI	opine (positive) → PI
purchase → PI	quit (negated) →PI

All classifiers and combinations of classifiers were evaluated against the manually annotated test set consisting of 2000 tweets. Two annotators were involved in the labelling. The agreement rate for the first round of an independent annotation has been 92.6%, and the ambiguous 7.4% were agreed on after discussion between the two annotators. In the final labelled test set, there are 537 PI and 1,463 non-PI labels. The upper bounds were obtained by training the supervised NBMT classifiers on the respectively represented test sets with 10-fold cross-validation. Note that the testing sets are not used for any of our split over-training models. OCA performances of all classifiers are given in Figure 4(a), and F-PI performances in Figure 4(b). For w_5, w_6, d_2 and d_3, a tweet may contain more than one GI keyword so that more than one local pattern might be extracted for one tweet. Similarly, Carlos'12 splits a tweet into sentences and then assigns each sentence with its intention category and sentiment and/or negation information. To keep the performances of different classifiers compatible, both OCA and F-PI were measured on the unit of tweets, i.e. a tweet will be classified as PI in the case of at least one of its constituent local patterns or sentences has been labeled with a PI tag; and classified as non-PI only if all its constituent local patterns or sentences have been labeled with non-PI tags.

Furthermore, as the task of PI/non-PI identification is an imbalanced classification in our corpus, i.e. there are much fewer PI texts than non-PI texts, we made a combination of classifiers (denoted as '*Combined*' in Figure 4) works in a way that a tweet should be labeled as PI if and only if all its combining classifiers have labeled the tweet as PI, and the tweet should be labeled as non-PI otherwise.

Evaluation results in Figure 4 show that our unsupervised split over-training approach achieved significant improvements over baseline classifiers, and the gaps between baselines performances and those of the upper bounds and the state-of-the-artclassifier: Carlos'12, were significantly narrowed. T-Test results in Table 7 show the significance of the improvements. For both OCA and F-PI, all 10-split, 5-split and combined PI identification models outperformed the baseline significantly, with p-values much smaller than 0.05. For single classifiers by 5-split over-training, the one on d_3 (3-level dependency local pattern) reports the best result of OCA 83.8% and F-PI 68.2%, with the OCA gap between its baseline and upper bound cut from 28.1% to 8.5%, and the F-PI gap from 18.7% to 7.1%; and for combined models, the combination of wt (whole tweets) and d_3 gives the best OCA of 84.6% cutting the gap to 7.7%, and the best F-PI of 70.4% cutting the gap to 4.9%.

Table 7: T-Test Results (Notes: For the test type, 1 means paired test and 3 means unpaired test.).

	Data1	Data2	Tails	Type	T-Test
OCA	Baseline	10-split	1	1	0.0004
	Baseline	5-split	1	1	0.0004
	Baseline	Combined	1	3	< 0.0001
	10-split	5-split	2	1	0.3262
F-PI	Baseline	10-split	1	1	0.0009
	Baseline	5-split	1	1	0.0012
	Baseline	Combined	1	3	< 0.0001
	10-split	5-split	2	1	0.3943

In general, our experiments showed that PI and non-PI tweets are more distinguishable locally around general intention words, while long distance information also captures some different classifying features. This explains why models on w_6 and d_3 were both improved when combined with wt.Furthermore, considering both OCA and F-PI measures, the difference between our upper bounds and Carlos'12 is not significant with the unpaired one-tail T-Test p-value of 0.73, which shows that our upper bounds are reasonable.

6. CONCLUSION

Our contributions in this chapter are three-fold: a) A formal definition of linguistically instantiated intentions, and especially GI and SI parts, which are the keys to identifying and distinguishing intentions; b) An unsupervised approach of split over-training for mining intentions, which may provide reasonable pre-processing for intention corpus labeling or intention knowledge acquisition; c) The empirically proved hypothesis that PI and non-PI texts are formally distinguishable.

We focused on the task of identifying Purchase Intention (PI) from social media, with a goal to solve the binary classification problem: whether a text expresses PI or non-PI. To the best of our knowledge, there is no large-scaled labeled social media intention corpus or linguistic knowledge database publicly available. Besides, the booming volume and wide variety of social media texts make the acquisition of linguistic knowledge greatly challenging, and large corpus labeling very expensive. Based on clustering and classification techniques, we proposed an unsupervised approach called split over-training for the purchase intention identification task, and simultaneously tried to answer the research question of how PI and non-PI social media texts are formally different from each other and related among themselves.

Experiments for PI identification were conducted on Twitter data represented on three levels: whole tweets, local patterns of word windows around general intention words, and local patterns of dependency structures around general intention words. F-PI and OCA were used as performance measure for over-trained split models and the output final models. Evaluation results on a manually labeled testing set of 2K tweets showed that our unsupervised method was effective and promising, with OCA and F-PI improved significantly over the baseline. Combined models of the split over-trained classifiers further reduced the gaps between respective baselines and upper bounds, while our upper bound classifiers are comparable to the state-of-the-art supervised method.

For future, we plan to extend our work to studies related to intention analysis in mental health domain e.g. for identifying user's intention to self-harm or suicide[31]. In order to facilitate adaptation of the proposed method to mental health domain, a few changes on the special intention keywords or phrases (SI) are needed. For instance, keywords like '*cut*', '*kill*', '*suicide*' provide clues of self-harm or suicide intention. Furthermore, we plan to leverage both structured and unstructured data for model learning[32]in order to achieve better intention identification performances.

ACKNOWLEDGEMENT

This paper has been supported by Center for Advanced Computing Technology (C-ACT), FakultiTeknologi Maklumat dan Komunikasi (FTMK), UniversitiTeknikal Malaysia Melaka, Malaysia.

REFERENCES

[1] C. S. Carlos and M. Yalamanchi, **Intention Analysis for Sales, Marketing and Customer Service**, in *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012), Demonstration Papers*, 2012, pp. 33–40.

[2] M. Bratman, ***Intention, plans, and practical reason***, 10th ed. Harvard University Press Cambridge, MA, 1987.

[3] T. M. Hiele, A. E. Widjaja, J. V. Chen, and T. Hariguna, **Investigating students' collaborative work to continue to use the social networking site**, *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 1.5 Special Issue, pp. 375–386, 2019.

[4] U. Rahardja, T. Hariguna, Q. Aini, and S. Santoso, **Understanding of behavioral intention use of mobile apps in transportation: An empirical study**, *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 1.5 Special Issue, pp. 258–263, 2019. https://doi.org/10.30534/ijatcse/2019/4581.52019

[5] M. Pontiki and H. Papageorgiou, **'There's No Way I Would Ever Buy Any Mp3 Player with a Measly 4gb of Storage': Mining In-tention Insights about Future Actions**, in *proceedings of the First International Conference on HCI in Business*, 2014, pp. 233–244.

[6] T. A. Han and L. M. Pereira, **State-of-the-art of intention recognition and its use in decision making**, *J. AI Commun.*, vol. 26, no. 2, pp. 237–246, 2013.

[7] W. W. Cohen, V. R. Carvalho, and T. Mitchell, **Learning to Classify Email into 'Speech Acts**,' in *Proceedings of the 2004 Conference on Empirical*

Methods in Natural Language Processing, 2004, vol. 4, pp. 309–316.

[8] M. Kröll and M. Strohmaier, **Analyzing human intentions in natural language text**, in *Proceedings of the fifth international conference on Knowledge capture - K-CAP '09*, 2009, p. 197. https://doi.org/10.1145/1597735.1597780

[9] D. J. Smith, **A Framework for Understanding the Training Process Leading to Elite Performance**,*Sport. Med.*, vol. 33, no. 15, pp. 1103–1126, 2003.

[10] D. Davidson, **Actions, reasons, and causes**, *J. Philos.*, vol. 60, no. 23, pp. 685–700, 1963. https://doi.org/10.2307/2023177

[11] G. E. M. Anscombe, ***Intention***. Harvard University Press, 2000.

[12] E. Charniak and R. P. Goldman, **A Bayesian model of plan recognition**, *J. Artif. Intell.*, vol. 64, no. 1, pp. 53–79, 1993.

[13] K. A. Tahboub, **Intelligent Human-Machine Interaction Based on Dynamic Bayesian Networks Probabilistic Intention Recognition**, *J. Intell. Robot. Syst.*, vol. 45, pp. 31–52, 2006.

[14] M. G. Armentano and A. Amandi, **Plan recognition for interface agents : State of the art**, *Artif. Intell. Rev.*, vol. 28, no. 2, pp. 131–162, 2007. https://doi.org/10.1007/s10462-009-9095-8

[15] V. Gupta, D. Varshney, H. Jhamtani, D. Kedia, and S. Karwa, **Identifying purchase intent from social posts**, in *Proceedings of the 8th International Conference on Weblogs and Social Media (ICWSM 2014)*, 2014, pp. 180–186.

[16] J. L. Austin, ***How to do things with words***. Oxford university press, 1975.

[17] J. R. Searle, **A Taxonomy of Illocutionary Acts.**, *Günderson, K. (ed.), Lang. Mind, Knowl.*, pp. 344–369, 1975.

[18] A. Qadir and E. Riloff, **Classifying Sentences as Speech Acts in Message Board Posts**, in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 748–758.

[19] J. Mildinhall and J. Noyes, **Toward a stochastic speech act model of email behavior.**, in *Proceedings of CEAS 2008*, 2008.

[20] R. Li, C. Lin, M. Collinson, X. Li, and G. Chen, **A dual-attention hierarchical recurrent neural network for dialogue act classification**, *CoNLL 2019 - 23rd Conf. Comput. Nat. Lang. Learn. Proc. Conf.*, pp. 383–392, 2019. https://doi.org/10.18653/v1/K19-1036

[21] P. Chandon, V. G. Morwitz, and W. J. Reinartz, **Do Intentions Really Predict Behavior? Self-Generated Validity Effects in Survey Research**,*J. Mark.*, vol. 69, no. 2, pp. 1–14, 2005.

[22] B. Schivinski and D. Dabrowski, **The Impact of Brand Communication on Brand Equity Dimensions and Brand Purchase Intention Through Facebook.**, *GUT FME Work. Pap. Ser. A. Gdansk Gdansk Univ. Technol. Fac. Manag. Econ.*, vol. 4, no. 4, pp. 1–24, 2013.

[23] D. P. Twitchell, M. Adkins, J. F. Nunamaker Jr., and J. K. Burgoon, **Using speech act theory to model conversations for automated classification and retrieval**, in *Proceedings of the 9th International Wo rking Conference on the Language-Action Perspective on Communicati on Modelling (LAP 2004)*, 2004, pp. 121–130.

[24] H. (Kathy) Dai, L. Zhao, Z. Nie, J.-R. Wen, L. Wang, and Y. Li, **Detecting online commercial intention (OCI)**, in *Proceedings of the 15th international conference on World Wide Web*, 2006, pp. 829–837.

[25] Y. He, C. Lin, and A. E. Cano, **Online sentiment and topic dynamics tracking over the streaming data**, in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, 2012, pp. 258–266.

[26] C. Lin, E. Ibeke, A. Wyner, and F. Guerin, **Sentiment--topic modeling in text mining**, *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 5, no. 5, pp. 246–254, 2015.

[27] A. T. Wibowo, A. Siddharthan, J. Masthoff, and C. Lin, **Incorporating constraints into matrix factorization for clothes package recommendation**, *UMAP 2018 - Proc. 26th Conf. User Model. Adapt. Pers.*, pp. 111–119, 2018.

[28] A. T. Wibowo, A. Siddharthan, C. Lin, and J. Masthoff, **Matrix factorization for package recommendations**, *CEUR Workshop Proc.*, vol. 1892, pp. 23–28, 2017.

[29] N. Banerjee, D. Chakraborty, A. Joshi, and S. Mittal, **Towards Analyzing Micro-Blogs for Detection and Classification of Real-Time Intentions.**, *Icwsm*, pp. 391–394, 2012.

[30] L. Kong, N. Schneider, S. Swayamdipta, A. Bhatia, C. Dyer, and N. A. Smith, **A Dependency Parser for Tweets**, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1001–1012. https://doi.org/10.3115/v1/D14-1108

[31] N. F. Abd Yusof, C. Lin, and F. Guerin, **Analysing the Causes of Depressed Mood from Depression Vulnerable Individuals**, in *Proceedings of the International Workshop on Digital Disease Detection using Social Media 2017 (DDDSM-2017)*, 2017, pp. 9--17.

[32] D. Liu and C. Lin, **Sherlock: A semi-automatic quiz generation system using linked data**, *CEUR Workshop Proc.*, vol. 1272, no. 1, pp. 9–12, 2014.