Customer Intentions Analysis of Twitter Based on Semantic Patterns

Mohamed Hamroun
SOIE Lab
University of Tunis
Tunisia
mohamed.hamrounn@gmail.com

Mohamed Salah Gouider
SOIE Lab
University of Tunis
Tunisia
ms.gouider@yahoo.fr

Lamjed Ben Said
SOIE Lab
University of Tunis
Tunisia
lamjed.bensaid@isg.rnu.tn

ABSTRACT

Intention analysis over Twitter offer organisations a fast and effective way to monitor the publics' desires towards their brand, business, directors,etc. A wide range of features and methods for user intention analysis have been researched in recent years. Due to the nature of micro blogs posts, mining user intentions is still a challenging task. In this paper, we introduce a novel method of using semantic patterns and ontologies. Our approach combines natural language syntax and semantics analysis to identify user intention. We conduct a case study of user intentions in the commercial field. Our experimental results show the importance and effectiveness of our proposed approach for detecting customer intentions.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*

General Terms

Theory

Keywords

Intention, semanitc, patterns, Twitter, Ontology, NLP, social networks

1. INTRODUCTION

Nowadays, social networks have become an important interaction media among worldwide users. Among the most used social networks is Twitter, a microblogging social network, with over 200 million users and about 400 million posts per day [13]. Today social media has been widely accepted as an active communication channel between companies and customers. Many companies regularly use social networking websites to promote new products and services, and post announcements to the customers. On the other hand, customers often post their comments to express their intentions towards products and services. The opportunity to capture user intention, sentiment has raised growing interest both within the scientific community, leading to many exciting open challenges, as well as in the business world, due to the remarkable benefits to be had from marketing and financial prediction. Intention analysis has received much attention among market research community as an effective approach for analyzing social media contents. Some highlighted applications of intention analysis include brand monitoring, campaign monitoring and competitive analysis. Syntactic approaches including word, manual templates to extracting user intentions have proven successful when applied to product and politic reviews that contain well-structured sentences [7] [14] [12] [6]. However, applying either approach to Twitter data faces several challenges. Firstly, tweets data are often composed of sentences of poor grammatical and syntactical structures due to the extensive use of abbreviations and irregular expressions in tweets [8]. Existing syntactic approaches to intention analysis mainly rely on parts of text in which intentions are explicitly expressed such as polarity terms, words, and their co-occurrence frequencies. However, intentions are often conveyed implicitly through latent semantics, which make purely syntactical approaches ineffective. In this paper, we propose a novel approach for automatically extracting semantic patterns for customer intentions analysis on Twitter. We refer for these patterns from now on as CI-Patterns. Our intent is to explore the possibilities of adding semantics to the patterns by using Semantic Web technologies. In our work, we benefit from the natural language processing steps performed by Wordnet, OpenNLP and the underlying OWL ontologies.

To be able to exploit the notion of the intention, it is necessary to formalize it according to a formal language. By Definition, Customer Intention is designed by three key components (Subject, Intention Verb, Object). Moreover, such Customer Intention is constrained in time and space. In this work, we focused on the three first elements of Customer Intention.

For evaluating and validating our approach, we apply 5 different datasets and compare the performance of CI-Patterns against methods trained from the state of the art. Our results report that our CI-Patterns almost outperform all our baseline methods, especially on twitter dataset by 3, 6% and 2% in precision and recall respectively. The main contribution of this paper consists of 3 parts:

- Propose a novel approach that automatically extracts patterns for customer intentions analysis of Twitter.
- Propose a new Representation of Customer Intentions ontology.
- Use extracted patterns as features for tweets pattern matching, and compare the performance against 3 state-of-the-art baselines on 5 different datasets.

2. RELATED WORKS

The principal context of our work is in the area of sentiment analysis, which is now a widely researched area because of the abundance of commentaries from weblogs, review sites, and social networking sites [2] [3]. On the one hand, researchers have studied and analyzed the user intent behind queries in web search. On the other hand [5] [4] [10] proposed classifying the intent expressed by web content creators and classified it as navigational or informational. The same authors published a follow-up study to bridge the gap between the link intent and the query intent, and how this gap filling will enhance web search quality. User intention has also been studied extensively in the commercial field [1] analyzed the relationship between search intent, result quality and searcher behavior in online purchases and how optimizing these interactions can enable more effective detection of searcher goals. Currently there are some researches related to our work. [6] introduced the novel task of identifying wishes. A wish corpus composed by political comments and product reviews was constructed and studied in details. A mix of manual templates and SVM-based text classifiers were applied on the wish corpus, and a method to identify more templates was also discussed. [12] interested two specific kinds of wishes: suggestions about existing products and intentions that indicate the author will buy a product. The paper limited their research to product reviews. They thought a majority of the suggestion wishes had pivotal phrases involving modal verbs such as "would", "could", "should" etc. So, rules based on modal verbs were extracted manual. [14] studied the problem of automatically identifying wishes in product reviews. These wishes are sentences in which the authors make suggestions about a product or show intentions to buy a product. So analyzing the wishes can insight into the minds of consumers to help product manufacturers, advertisers, and others looking to discover what really customer needs. This paper firstly proposed a keyword strategy to find candidate wish sentences, then sequential pattern are mined from these sentences labeled by manual, lastly using the patterns as features, a classifier is trained to identify wish sentences in product reviews.

3. PROPOSED APPROACH

Based on [9], we propose a lexico semantic patterns based approach which exploits domain ontology for patterns creation. These patterns are applied on commercial tweets to extract customer intentions. The pattern approach proposed in this section differs with respect to several aspects from approaches presented above since they do not capture the semantic context of text (mainly tweets), while our approach aims for the semantic description of the context. Our intent is to explore the possibilities of adding semantics to the patterns by using Semantic Web technologies. In our work, we benefit from the natural language processing steps performed by Wordnet [11], OpenNLP and the underlying OWL ontologies. In our case, ontologies will be used for retrieving relevant customer intentions in a semantically enhanced way. In addition discovered knowledge is stored in separate ontological database which might be updated with new facts. The creation of ontology is generated in parallel with the execution of different stages(subsection2). As shown in Figure 2, our framework consists of two main parts:

(a) A Preprocessing stage: before can be employed to match patterns in tweet feeds. A list of processing steps needs to be applied such as tokenization, remove redundant letters in words, stemming, and Part of Speech tags (OpenNLP POSTagger) which are dealt with by the OpenNLP framework.

(b) Patterns Engine part consists of two core components

- Lexico semantic pattern induction,
- Matching patterns to tweets.

3.1 Patterns definition

For CI-Patterns definition, we benefit from the research that has been proposed by [9]. Where each pattern is described by a left hand side (LHS) and right hand side (RHS). Once, the RHS is matched to tweets to be processed. In our case, the LHS describes a new representation for intention that defines a relation between three key components a subject (sub) and object (obj) using a predicate (I):

- A Subject $sub that can be a person, organizarion.
- An Intention: a verb predicate denoted by $I, and refers to an object $obj.
- An Object denoted by $obj that can be product, a product component, or service.
- $obj is described with a pair (T,W) where T is hierarchy of parts, sub parts (Product) and W is a set of attributes (Product components). Each part or sub part also its own of attributes.

Table 1: Exemple of Intention components

Tweet	@Jhon intends to buy an #apple #watch
Components	• $sub:(intendee) jhon instance of Class Person, • $I (has intention): intends to buy, • $obj: (intended object)an apple watch instance of Class Product.

We denote the LHS of a pattern as follows:

$$(\$sub, \$I, \$obj) : RHS \tag{1}$$

The subject, relation, and object described in the LHS need to be identified in the RHS in order to provide a link between tweet and a new extracted fact. This can be done using labels, which are represented as words preceded by a "$" and followed by a colon and an equality sign, as well as a description of the attached token. Whenever the RHS matches with a sentence, the tokens with associated labels are filled in the LHS of the pattern.

($sub, kb: intends, $obj):- $sub:=kb: Jhon $obj:=kb:apple watch (2)

Note that "$I:" represents an intentional verb, which in our case refers to the ontology. The RHS on the right hand describes a pattern that has to be identified in tweets. We define a pattern as an ordered collection of tokens that are divided by spaces. Based on OpenNLP POSTagger tool, our approach supports a set of syntactic categories to describe the lexical category of the token. This step is defined a preprocessing stage. As shown in the Table 2, We distinguish between various verbs, nouns, adjectives, prepositions, coordinating conjunctions (e.g., "as well as") and cardinal numbers.

Table 2: Common lexical categories

Category	Description
CC	Coordinating conjunction
D	Determiner
IN	Preposition
JJ	Adjective
NN	Noun
NNP	Proper Noun
PP	Pronoun
RP	Adverb
VB	Verb, base form
VBZ	Verb, 3rd person singular present

3.2 Customer Intentions Ontology

This section illustrates the creation of the domain ontology and its employment to enhance the patterns induction process. As already mentioned, the basic idea behind the proposed approach is to take advantage of domain ontology for enhancing the pattern learning process regarding the knowledge contained in commercial tweets. When ontologies are employed in the patterns, potentially one pattern can describe multiple representations. For employing ontology in patterns, our methodology consists of two phases: (a) generating of the ontology representation (b) employing ontology representation in patterns learning.

3.2.1 *Generating of ontology representation*

The knowledge representation of our intentional ontology that we use is as follows. The conception of ontology is generated in parallel of the execution of the preprocessing step. Our ontological representation is a simple list of key words denoted by intentional verbs representing the relation between concepts. The preprocessing step, including the tokenization, stemming and POS tagging, allows making a syntactic analysis on each token of the tweet. Starting form syntactic tags, we create a semantic representation. The later uses the semantic lexical database Wordnet[11] to find the synonyms and hyponyms of verbs, which retrieves synsets (group of synonyms words or collections) of the synonyms and hyponyms of every given word. Each Synonym is then added to the ontology and associated with initial verb. Syntactically, in the OWL representation of the ontology associations are expressed via the OWL: equivanlentClass and OWL:SubClassof constructed respectively. According to our pattern definition, pattern is composed of three key components subject, predicate, and object. As cited previously

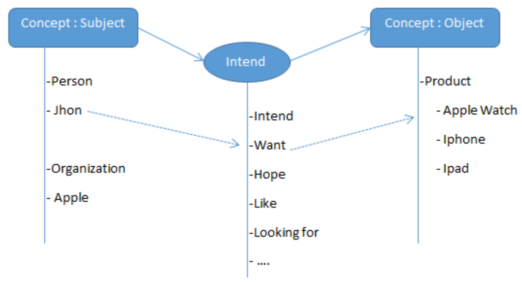

Figure 1: Ontology visualization

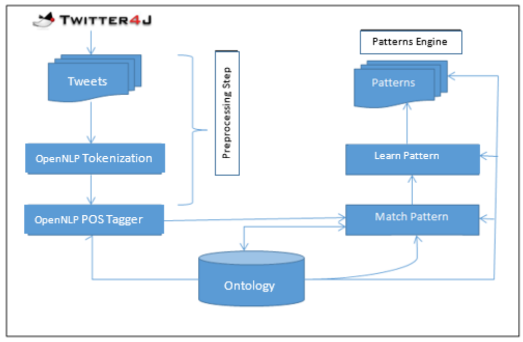

Figure 2: Proposed Architecture

the intentional verb is describing the relation (predicate) between the two concepts subject and object. For example, if we take the verb "intend": this verb makes the link between the concept subject that can be an instance of Class person or an instance of Class organization and the concept object that can be an instance of a Class Product or an instance of a Class Service. The main class is "intend" and sub classes of this verb are such as: want, hope, looking for, etc.

3.2.2 *Employing Customer Intentions Ontology in Patterns*

By employing ontology elements, we are adding semantic to patterns. As stated earlier, the LHS of the pattern is used for recognizing concepts, and it is a triple that describes the relationship between a subject and an object. In our case, concepts are groups that share the same proprieties. By using labels, we can refer in the LHS to a concept found on the RHS. For instance, a pattern such as

($sub, kb:hasIntention, $obj):-
$sub:=[kb:Person]
kb:look for
$obj:=[kb:Product]

. If we apply this pattern to the following tweet,

"*I am looking for a #brand #new #car to replace my old Ford Focus*". It would identify an intention of buying a new car. This information can then be used to update the ontology and to extract new facts.

4. EXPERIMENTAL SETUP

This section evaluates the effectiveness of our model and discusses the results. We first present Datasets and selected evaluation measures to perform the evaluation setup in subsection1, followed by the results in subsection 2.

4.1 Evaluation Settings

4.1.1 Datasets

For evaluation the performance of the executed CI-patterns, we use 4 datasets used in our baseline methods listed in the state of arts. These primarily belonged to the domains of electronics and retail banking collected from different sources such as popular consumer review sites (such as Epinions.com and MouthShut.com). Of these, we chose reviews about the Apple iPod (Data1), Digital Camera (Data2), TV (Data3) and a collection of banking reviews about five leading US banks (Data4). We also used TREC Microblog Dataset 2011. The dataset contains 16 million tweets, with about 5 million English tweets, collected from January 23, 2011 to February 08, 2011. TREC 2011 provides 50 topics manually picked by human assessors of the National Institute of Standards and Technology (NIST). This dataset is reliable and stable to evaluate the model. The sizes of the datasets are summarized in Table 5. The ontologies employed in our experiments contain major domain concepts. Our ontology contains a small of commonly used, well known, commercial concepts and intentional verbs. Example of ontology concepts are: as subject (i) (person, company, organization) and as object (ii) (product, product component) TableII describes different relation extracted from the TREC Dataset.

Table 3: Relations for the commercial domain, used for evaluation purposes

Subject	Relation(Intentional verb)	Object
Person	want	Product
Person	Has wish	Product
Person	Look for	Product

4.1.2 Evaluation measures

In our experiments, we compare the performance of our CI-patterns and the performance of our baseline methods in terms of precision, recall and F1 measure. These measures are suitable because our objective is to identify intention posts. These measurements are defined as follows:

$$P = \frac{|Relevant \bigcap Found|}{|Found|}, \tag{3}$$

$$R = \frac{|Relevant \bigcap Found|}{|Relevant|} \tag{4}$$

Where Relevant is the set of relevant intention posts and Found is the set of found intention posts. There is a trade-off between precision and recall, and hence we compute the F1 measure. The F1 measure is applied to compute an even combination, i.e., the harmonic mean of precision and recall:

$$R = \frac{2 \times P \times R}{P + R} \tag{5}$$

4.2 Experimental Results

In this section, we now compare CI-patterns with the baselines described in the state of the art. Experiment tasks are done using 10 fold cross validation. The first task in our evaluation aims to compare the performance of CI-patterns against the performance of baselines in terms of precision (P) and recall (R). Table 5 shows the results across all datasets. The highest precision of CI-patterns is achieved on the TREC Dataset with 55, 59% which explains the significance of improvement. While, the highest recall of 72% is obtained on the Data1 dataset. On the other hand, the lowest performance in precision is obtained using Wu method on the Data2. Also, Glodberg method produces the lowest recall of 24, 9% on the TREC Dataset. On average, CI-patterns outperforms well on the TREC Dataset by 3, 6% and 2% in precision and recall respectively.

Table 4 represents the number of CI-Patterns extracted from all datasets and used as features for the classification setup.

Table 4: Number of CI-Patterns extracted from all datasets

DataSets	Data1	Data2	Data3	Data4	TREC
# CI-Patterns	18	20	23	22	24

At the second, we focus on the F1 measurement. Figure 3 reports results obtained in F1 score using equation5 across all datasets.

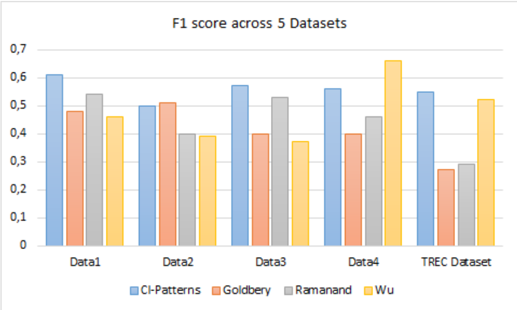

Figure 3: F1 measure results of CI-Patterns against baseline methods

Table 5: precision and recall results across 5 different datasets

Datasets	Total of Posts	#Intentions	Goldbery		Ramanand		Wu		CI-Patterns	
			P	R	P	R	P	R	P	R
Data1	21147	90	63%	40%	58,81%	50%	47%	46,3%	54,3%	72%
Data2	6850	224	45%	60%	30,67%	60%	30%	58%	58%	45%
Data3	1236	355	38%	42%	80%	40%	39%	44%	80%	45%
Data4	2289	28	49,2%	35%	57,14%	39%	79%	57%	57%	56,8%
TREC DataSet	5000	3200	33%	24,9%	35%	25%	52%	53,59%	55,59%	55,28%

5. CONCLUSION

This work addresses the extraction of semantic patterns for customer intentions analysis of Twitter. Our approach does not rely on manual or limit syntactic templates of intentions detection however it employs ontology concepts and relations designed by intentional verbs. We applied our approach on 5 different datasets, including product review and commercial Twitter posts, and validated the extracted patterns by using them as features for the classification task. We used 3 baseline methods to evaluate the performance of our approach. Our proposed approach CI-Patterns showed a consistent and superior performance over baseline methods especially on the TREC Dataset.

6. REFERENCES

[1] A. Ashkan, C. L. Clarke, E. Agichtein, and Q. Guo. Classifying and characterizing query intent. In *Advances in Information Retrieval*, pages 578–586. Springer, 2009.

[2] E. Cambria, B. Schuller, B. Liu, H. Wang, and C. Havasi. Knowledge-based approaches to concept-level sentiment analysis. *IEEE Intelligent Systems*, (2):12–14, 2013.

[3] E. Cambria, H. Wang, and B. White. Guest editorial: Big social data analysis. *Knowledge-Based Systems*, (69):1–2, 2014.

[4] N. Dai, X. Qi, and B. D. Davison. Bridging link and query intent to enhance web search. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, pages 17–26. ACM, 2011.

[5] N. Dai, X. Qi, and B. D. Davison. Enhancing web search with entity intent. In *Proceedings of the 20th international conference companion on World wide web*, pages 29–30. ACM, 2011.

[6] A. B. Goldberg, N. Fillmore, D. Andrzejewski, Z. Xu, B. Gibson, and X. Zhu. 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*, pages 263–271. Association for Computational Linguistics, 2009.

[7] H. Gómez-Adorno, D. Pinto, M. Montes, G. Sidorov, and R. Alfaro. Content and style features for automatic detection of usersâĂŹ intentions in tweets. In *Advances in Artificial Intelligence–IBERAMIA 2014*, pages 120–128. Springer, 2014.

[8] B. Hollerit, M. Kröll, and M. Strohmaier. Towards linking buyers and sellers: detecting commercial intent on twitter. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 629–632. International World Wide Web Conferences Steering Committee, 2013.

[9] W. IJntema, J. Sangers, F. Hogenboom, and F. Frasincar. A lexico-semantic pattern language for learning ontology instances from text. *Web Semantics: Science, Services and Agents on the World Wide Web*, 15:37–50, 2012.

[10] C.-H. L. Lee and A. Liu. Modeling the query intention with goals. In *Advanced Information Networking and Applications, 2005. AINA 2005. 19th International Conference on*, volume 2, pages 535–540. IEEE, 2005.

[11] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller. Introduction to wordnet: An on-line lexical database*. *International journal of lexicography*, 3(4):235–244, 1990.

[12] J. Ramanand, K. Bhavsar, and N. Pedanekar. 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*, pages 54–61. Association for Computational Linguistics, 2010.

[13] K. Wickre. Celebrating #twitter7. twitter blog, 2013.

[14] X. Wu and Z. He. Identifying wish sentence in product reviews. *Journal of Computational Information Systems*, 7(5):1607–1613, 2011.