

Towards Linking Buyers and Sellers: Detecting Commercial Intent on Twitter

Bernd Hollerit

Graz University of Technology
Inffeldgasse 13
A-8010 Graz, Austria

bernd.hollerit@student.tugraz.at

Mark Kröll

Know-Center, Division for Knowledge
Relationship Discovery
Inffeldgasse 13
A-8010 Graz, Austria

mkröll@tugraz.at

Markus Strohmaier

Knowledge Management Institute,
Graz University of Technology and
Know-Center, Inffeldgasse 13
A-8010 Graz, Austria

markus.strohmaier@tugraz.at

ABSTRACT

Since more and more people use the micro-blogging platform Twitter to convey their needs and desires, it has become a particularly interesting medium for the task of identifying commercial activities. Potential buyers and sellers can be contacted directly thereby opening up novel perspectives and economic possibilities. By detecting commercial intent in tweets, this work is considered a first step to bring together buyers and sellers. In this work, we present an automatic method for detecting commercial intent in tweets where we achieve reasonable precision 57% and recall 77% scores. In addition, we provide insights into the nature and characteristics of tweets exhibiting commercial intent thereby contributing to our understanding of how people express commercial activities on Twitter.

Categories and Subject Descriptors

I.2.6 [Artificial Intelligence]: Learning knowledge acquisition;

I.2.7 [Artificial Intelligence]: Natural Language Processing

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Commercial Intent, Twitter, Knowledge Acquisition

1. MOTIVATION

Micro-blogging platforms such as Twitter provide suitable means to distribute personal messages thereby opening up unprecedented economic opportunities. People often tweet about their needs and desires. They also tweet about things they want to get rid of. From an economic perspective, it would be valuable to, e.g., provide respective product information or purchase interests. As a prerequisite to “link buyers and sellers”, we need to detect tweets containing commercial intent. In this work, we provide insights into the nature of these tweets and present an automatic method. We leave the task to link buyers and sellers to future work.

Making use of and understanding Twitter content has been an ongoing endeavor over the past years including tasks such as extracting relevant and interesting key phrases (cf. [14]),

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author's site if the Material is used in electronic media.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.

ACM 978-1-4503-2038-2/13/05.

detecting real-time events ([12]) and analyzing sentiment expressions or opinions (cf. [10], [7]). Analyzing (commercial) intent is orthogonal to sentiment analysis as well as opinion mining and thus provides a different - an *intentional* perspective. Intent analysis (cf. [8]) also deals with a different temporal focus than sentiment analysis where a present (emotional) state is approximated.

This work was inspired by research conducted in the area of query log analysis. Dai et al. [4] were among the first to identify search queries exhibiting “online commercial intention”, i.e. where users intend to commit a commercial activity such as auction, paid service or purchase. Since queries do not contain much information on their own, i.e. 2-3 tokens on average, they characterized queries by including content of search results as well as landing pages. Follow up research was conducted by Ashkan et al. [1], yet their focus was different, i.e. sponsored search. Detecting commercial intent in queries partly served as a pre-processing step to analyze the correlation of click-through behavior and rank/location of sponsored links or ads. Guo et al. [5] attempted to differentiate between search intents by using interaction features such as mouse movements or scrolling behavior. In previous work Strohmaier et al. [13] showed that search query logs represented a viable, yet largely untapped, source for acquiring knowledge about human goals.

In the remainder of this paper we will define commercial intent in tweets and provide rationales for our definition. We will provide detailed statistics on 1335 annotated tweets which we obtained from Twapperkeeper¹, a service platform for collecting, storing and exporting tweet archives. We then used WEKA², a Java-based machine learning toolkit, to apply feature engineering and to learn a classification model. Using word and part-of-speech n-grams as attributes, we achieve precision scores of up to 57% and recall scores of up to 77% by applying 10-fold cross-validation. In addition to economic opportunities, our work could (i) improve spam detection (cf. [2]) for Twitter messages to prevent unsolicited/unrelated offers or advertising and (ii) introduce an additional search facet on Twitter.

2. COMMERCIAL INTENT IN TWEETS

We define a tweet exhibiting commercial intent (CI) whenever

A tweet (1) contains at least one verb and (2) describes the user's intention to commit a commercial activity (cf. [4]) (3) in a recognizable way (cf. [6]).

¹ <http://twapperkeeper.com> (now integrated into HootSuite Archives)

² <http://www.cs.waikato.ac.nz/ml/weka/>

The first part of the definition addresses the crucial role verbs play in explicating commercial intent in textual resources. An explicit representation saves us from having to deal with ambiguous expressions. “Recognizable” refers to what Kirsh ([6]) defines as “trivial to identify” by a subject within a given attention span. By “trivial to identify” Kirsh means the ability to make a decision in constant time. This definition was adapted from previous work ([13]) to serve the specific needs of our research.

We then used Twapperkeeper to generate tweet archives - an archive admitted only tweets containing a particular keyword. When compiling our list of keywords, we included keywords used for detecting commercial intent in search query logs (cf. [4]) such as price or discount, adjectives such as cheap and additional verbs which we selected from Levin’s ([9]) verb classes 13.1, 13.5.1 and 13.5.2 such as to trade. Focusing on a list of commercial keywords may appear restrictive in a sense that certain constructions or other potential keywords are excluded. However, while annotating tweets we experienced that randomly chosen tweets had little chance of exhibiting commercial intent at all. For the purpose of generating training data, we thus took only tweets into consideration which contained at least one commercial keyword. Some of these, e.g. coupon, discount or lease were immediately discarded because they didn’t provide useful or enough samples. By carefully examining the samples, we decided on following 16 keywords to use for the annotation process:

advertise, auction, bidding, buy, cheap, cost, deal, find, get, market, price, purchase, rent, retail, sale, sell

Per keyword we collected and annotated 100 tweets. Some of the Twapperkeeper archives didn’t contain 100 tweets including following keywords: cost (21), deal (87), get (84), market (22), price (32), rent (13) and sale (76). In total, we annotated 1335 tweets with respect to commercial intent.

During the process of annotation, we were not only interested in whether a tweet contained commercial intent or not but also (i) whether the intent was implicit or explicit and (ii) whether the commercial activity encompassed buying or selling intention. To give some clarifying examples:

Explicit vs. Implicit: The tweet “Facing Repossession, Let us buy your house for cash now <http://tiny.ly/G7Rw>” explicitly expresses the intent to buy a house. In contrast, the tweet “Debating on buying the pair of 80s cop shades...” contains to a certain extent commercial intent, but it is not explicitly stated rather as a possibility in the future. In our understanding, being aware of this type of commercial intent has also economic value.

Buying vs. Selling: If it contains commercial intent, does the person want to buy or sell something? “I’ll buy the Joe-Nastics dvd” (buy intent) vs. “I’m selling my black emperor scorpion” (sell intent).

We only consider keywords which indicated commercial intent at least once during annotation leaving us with 8 keywords, i.e. auction, bidding, buy, cheap, find, purchase, retail and sell. In general, we observe a low density of commercial intent even with commercially indicative keywords present. In total, 120 annotated tweets contained commercial intent - 70 of them exhibited buying intent, 50 selling intent - 39 of them exhibited explicit intent, 81 implicit intent.

Figure 1 gives an overview of every keyword’s share in the number of tweets containing commercial intent and thus also

reflects its value. To give an example, out of 120 tweets containing commercial intent, there are 34 tweets containing the keyword buy, hence 28%.

The distribution in Figure 1 also confirms a natural assumption that keywords such as buy, sell or cheap are inherently connected with commercial activities. A fitting anecdote at this point is that originally the keyword list also contained additional adjectives such as “expensive”, yet it did not turn out to be a good indicator.

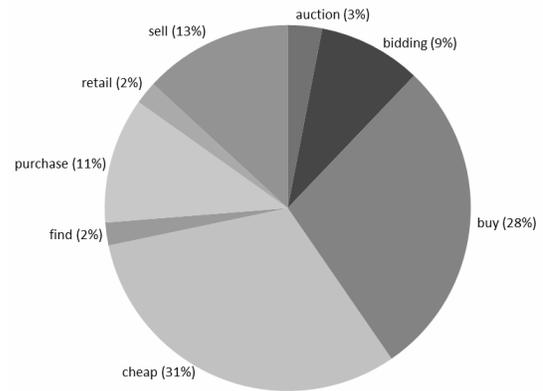


Figure 1: Each keywords share in annotated tweets containing commercial intent (in total 120).

Figure 2 and Figure 3 provide a detailed picture of commercial intent amongst annotated tweets further characterized by implicit vs. explicit intent and buying vs. selling intent for each keyword.

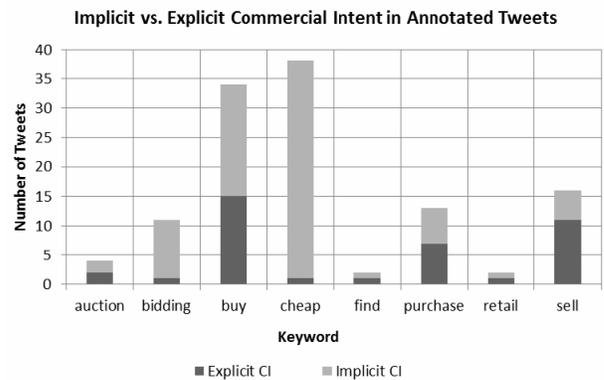


Figure 2: Distribution of commercial intent per keyword further characterized by contrasting implicit and explicit intent.

Figure 2 shows that most of the commercial intent is implicit in nature with the keyword cheap as main contributor. That suggests that a product first has to be or has to become cheap to commercially act. Explicit commercial intent is prominent with respect to the keyword sell indicating that people are more explicit (and thus certain) when it comes to selling something.

As expected, Figure 3 corroborates that keywords buy and purchase are good indicators for buying intent and the keyword sell for selling intent. It also suggests that there is more buying than selling intent on Twitter. Besides the keyword sell, selling intent is prominent with the keyword cheap indicating that product offerings often go hand in hand with “low in price” announcements.

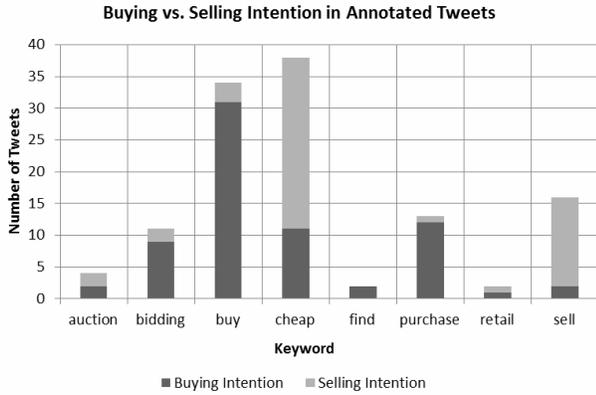


Figure 3: Distribution of commercial intent per keyword further characterized by contrasting buying and selling intent.

In brief, the keywords buy and cheap appear to be good indicators for commercial intent since they account for ~60%. The keywords auction, find and retail on the other hand only provide us with 0% - 4% of commercial intent.

3. AN AUTOMATIC METHOD

We devise a classification approach that aims to perform the task of classifying tweets into one of the two categories (containing/not containing commercial intent) automatically. For the class “containing commercial intent” we used annotated tweets containing implicit as well as explicit commercial intent. As attribute types we use word and part-of-speech³ n-grams. Extensive pre-processing and filtering was necessary to apply the Stanford part-of-speech tagger⁴ and the WEKA machine learning toolkit. The filtering included following steps: (i) remove all characters except for A-Z, a-z, 0-9 and spaces, (ii) make every character lowercase, (iii) if absent, append a period to every tweet and (iv) replace two or more consecutive spaces by exactly one space. For generating the word attributes, we then applied WEKA’s pre-processing suite which included token stemming and n-gram creation setting the n-gram parameter to a range from two to five (other parameters for n did not improve the results).

3.1 Discriminative Attributes

We used WEKA to compute the value of the chi-squared statistic with respect to the class to obtain a ranking of the most valuable attributes in the classification task at hand. Table 1 shows the top 20 most discriminative attributes. To aid readability, descriptions of selected part-of-speech tags are provided according to the Penn Treebank Tag Set: VB represents the base form of a verb, NN represents a noun in singular form, JJ represents an adjective, FW represents a foreign word, MD represents a modal and DT represents a determiner. Textual attributes occupy the top four spots in the ranking. In general, some of these textual attributes indicate commercial intent more openly than others. From our observations, the attribute “buy cheap” is often used as an invitation. The attribute phrase “i want to” expresses intent on a general level and therefore appears to be a good indicator for commercial intent as well since it is often followed by a statement about an economic product.

³ <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/COP-HTMLDemo/PennTreebankTS.html>

⁴ <http://nlp.stanford.edu/software/tagger.shtml>

Table 1: The top 20 most discriminative attributes are illustrated resulting from applying WEKA’s chi-square attribute selection.

Rank	Attribute	Example Tweets
1	buy cheap	'buy cheap alberto vo5 shampoo strawberries'
2	to buy	'np pink fridayi think im going to buy it tomorrow'
3	for sale	'for sale apple iphone 4g 32g/apple iphone 3gs 32gb buy 2 get 1 free'
4	check out	'#quilt lovers, check out @heyporkchop's flea market fancy scraps for auction'
5	VB DT JJ	'dear allstarweekend please come back to michigan so we can buy those new shirts d',RB RB VB VB RB TO VB IN PRP MD VB DT JJ NNS LS'
6	VB JJ NN CD	'buy cheap braun 5270 silkpil x'
7	NN NN CD CD	'classifieds i am selling my gmc envoy xl 2003 for gooddemand sr 35 000 slightly negotiablei am the secon httpbitly3g1e4'
8	have to buy	'cooking carbonnade and for drink just wine ... i have to buy food tomorrow :S'
9	low price	'buy cheap blue banana dresses low price everyday @amazon.co.uk http://amzn.to/9hzjhq'
10	NN NN JJ CD	'buy cheap 25 usb 20 to sata hard drive hdd aluminum external 25 usb 20 to sata hard drive hdd aluminum e httpbitlybzstip'
11	NN NN CD CD NN	'for sale apple iphone 4g 32gapple iphone 3gs 32gb buy 2 get 1 free httpbitlyhcazwp'
12	i want to	'delhi buy sell i want to sell my nokia n97 i want to sell my nokia n97 which is new brand phone with all f httpbitlyb7kgl'
13	NN IN DT JJS	'buy your new or used bmw in ebay for the best possible price more info httpbitlybf0zqr'
14	VB DT JJ NN	'about to go to first Friday with codynotontor to find an xmas present for dianevicars anyone want to join free booze and cheese'
15	VB JJ NN CD NN	'xbox 360 system link cable buy cheap xbox 360 system link cable buy low price from here now consider yourself con httpbitly9wiiej'
16	NN JJ CD	'buy cheap bikemaster turn signal honda rear 251036 for 1756'
17	JJ NN NN FW	'billion is milyard rt ivnsari i need 5 billion to buy an iphone but i just have 1 billion hwhw'
18	FW MD	'i think i may need to buy one more coat'
19	VB JJ NN	'i nid to buy stuffwoahlong listiono where 2 start 4rumsmh'
20	im selling	'im selling @djtazach dj equipment and the bidding starts at?!?! lol http://plixi.com/p/43638541'

Attributes such as “to buy” or “im selling” are obvious indicators for commercial intent. We also noticed that lower ranked textual attributes include expressions of general intent such as “i really need” and often combine these expressions with commercial keywords such as “i want to sell” or “want to buy”. Part-of-speech n-grams appear to be good indicators as. Yet, while some of them do match commercial phrases, the other ones do not but rather appear to often co-occur with textual keywords such as sell or buy.

3.2 Learning a Classification Model

WEKA provides us with a range of different classifiers which allow us to compare classifiers, e.g. with respect to linear vs. non-linear decision boundaries. For our experiments, we used annotated tweets exhibiting implicit as well as explicit commercial intent for the positive class, i.e. 120 tweets. The remaining 1215 tweets were representatives for the negative class. As attributes we used word as well as part-of-speech n-grams. Precision and recall scores were calculated by means of 10-fold cross-validation. Since we are mainly interested in achieving high values for the positive class, i.e. tweets containing commercial intent, we only report precision, and recall scores for the positive class.

For the classification task, we applied several classification models capable of generating linear as well as non-linear decision boundaries including a linear Support Vector Machine and Nearest Neighbor algorithms. Best recall scores of 77.4% were achieved using a Bayes Complement Naïve Bayes classifier, a classification model which attempts to address the shortcomings of the Naïve Bayes classifier (cf. [11]) in the textual domain. Best precision scores of 57.1% were achieved by using a linear logistic regression classifier. Based on the antagonistic nature of the two metrics, only one of them could be optimized leading to bad F1 scores in either case. We understand these moderate scores as a first baseline for further investigations. Examining classified tweets shows that incorrect part-of-speech tagging often leads to false negatives, e.g. in the tweet “*just bidding on a bloody hamster cage whatamidoing*” where the gerund “bidding” was annotated as a noun. Incorrect part-of-speech tags are most likely due to the difference in syntactical and grammatical structure of tweets versus the Wall Street Journal corpus, the tagger has been trained upon. The example also reveals that a better understanding of language usage on Twitter - which often is informal and colloquial - might be advantageous in the process.

4. DISCUSSION AND CONCLUSION

This work addresses the detection of commercial intent on Twitter and thereby contributes to opening up economic possibilities by being capable (i) of contacting potential buyers or sellers directly, (ii) of providing more targeted ad banners as well as preventing unsolicited/unrelated offers or advertising or (iii) of recognizing product trends and analyze them over time. These possibilities might eventually bring buyers and sellers closer together.

We annotated a set of 1335 tweets gaining insights into the nature of commercial intent, i.e. which keywords are indicative for commercial activity. We applied WEKA to learn a classification model. Using word and part-of-speech n-grams as attributes, we achieve precision scores of up to 57.1% and recall scores of up to 77.4% by applying 10-fold cross validation. We understand these scores as first baseline for our ongoing endeavor to improve the intent detection method. Our next steps thus include (i) calculating an inter-rater agreement κ (cf. [3]) to validate our definition of commercial intent (ii) investigating further attribute types such as temporal or retweet information, (iii) analyzing present URLs including the respective site’s content and (iv) identifying potential commercial keywords in an automated way. In addition, our analysis suggests that identifying tweets which express intent in general might be a valuable pre-processing step for detecting commercial intent.

Besides sketching potential economic perspectives, our work introduces a novel, an *intentional* dimension to characterize textual content on Twitter and could thus add an additional search facet on Twitter.

5. ACKNOWLEDGMENTS

The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

6. REFERENCES

- [1] Ashkan, A. and Clarke, C. 2009. Term-based commercial intent analysis. In Proc. of the International Conference on Research and Development in Information Retrieval.
- [2] Benczúr, A., Bró, I., Csalogány, K. and Sarlós, T. 2007. Web spam detection via commercial intent analysis. In Proceedings of the 3rd International Workshop on Adversarial Information Retrieval on the Web.
- [3] Cohen, J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement.
- [4] Dai, H., Zhao, L., Nie, Z., Wen, J., Wang, L. and Li, Y. 2006. Detecting online commercial intention (OCI). In Proceedings of the World Wide Web Conference.
- [5] Guo, Q. and Agichtein, E. 2010. Ready to buy or just browsing?: Detecting web searcher goals from interaction data. In Proceedings of the International Conference on Research and Development in Information Retrieval.
- [6] Kirsh, D. 1990. When is information explicitly represented? Information, Language and Cognition – The Vancouver Studies in Cognitive Science.
- [7] Kouloumpis, E., Wilson, T. and Moore, J. 2011. Twitter sentiment analysis: The good the bad the OMG! In Proc. of the International Conference on Weblogs and Social Media.
- [8] Kröll, M. and Strohmaier, M. 2009. Analyzing human intentions in natural language text. In Proceedings of the International Conference on Knowledge Capture.
- [9] Levin, B. 1993. English verb classes and alternations: A preliminary investigation. University of Chicago Press.
- [10] Maynard, D. and Funk, A. 2011. Automatic detection of political opinions in tweets. In Proceedings of the 1st Workshop on Making Sense of Microposts at ESWC’11.
- [11] Rennie, J., Shih L., Teevan J. and Karger, D. 2003. Tackling the poor assumptions of Naive Bayes text classifiers. In Proc. of the International Conference on Machine Learning.
- [12] Sakaki, T., Okazaki, M. and Matsuo Y. 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. In Proceedings of the World Wide Web Conference.
- [13] Strohmaier, M. and Kröll, M. 2012. Acquiring knowledge about human goals from search query logs. Information Processing and Management 48, 1.
- [14] Zhao, W., Jiang, J., He, J., Song, Y., Achananuparp, P., Lim, E. and Li, X. 2011. Topical key phrase extraction from Twitter. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: (HLT ’11).