# Linking Buyers and Sellers:

## Detecting Commercial Intent on Twitter

Bernd Hollerit, Mark Kröll and Markus Strohmaier
Graz University of Technology, Austria & Know Center

#MSM 2013:
Making Sense of Microposts

@ WWW 2013
Rio de Janeiro, Brazil

Bernd Hollerit
bernd.hollerit@student.tugraz.at

# Schedule

- Motivation
- Related work
- Definition of Commercial Intent
- Collecting and Annotating Tweets
  - Different Kinds of CI
    - Annotated Tweet Archives
    - Percentage of CI per Tweet Archive
    - Buying vs. Selling Intention Percentage
    - Explicit vs. Implicit CI Percentage
- Experimental Setup
  - Part-of-Speech Tagger
  - Feature Generation
    - WEKA Top 10 Most discriminative attributes
  - Learning a Classification Model
    - Classification Model Preliminary Results
- Conclusion
- Selected Comments from Reviewers
- References
- Q & A

# Motivation

- People are prolific in writing their desires and needs on Twitter, e.g. buying and selling products
- Vision: bringing those groups together
  - First step would be to detect commerical intent in tweets
- Opens up economic opportunities
  - People can be contacted directly on Twitter

# Related work

- Analyzing Intent
  - Search query logs [Dai06], [Ashkan09], [Guo10]
  - Textual resources [Kröll09]

- Exploiting data on Twitter to…
  - detect real-time events [Sakaki10]
  - to extract relevant and interesting key phrases [Zhao11]
  - to analyze sentiment expressions or opinions [Maynard11], [Kouloumpis11]

# Definition of Commercial Intent (CI)

- *The tweet (1) contains at least one verb, (2) describes the user's intention to commit a commercial activity* (cf. [Dai06]) *and (3) in a recognizable way* (cf. [Kirsh90]).

- (1) tweets without verbs like „house" or „expensive car" don't convey intent explicitly
- (2) commercial activities encompass buying, selling or bargaining intent
- (3) „Recognizable" refers to what Kirsh defines as „trivial to identify" by a subject within a given attention span.
  - the ability to make a decision in constant time.

# Different Kinds of Commercial Intent

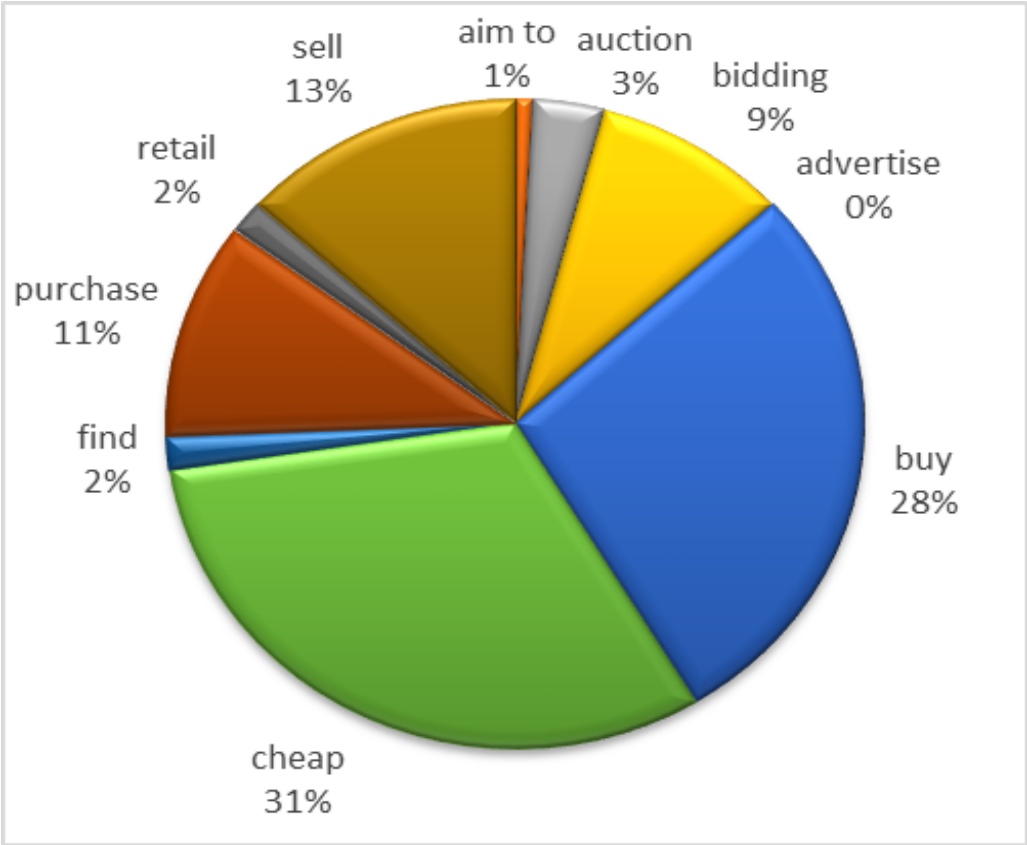
- Buying vs. Selling Intent
  - If the tweet contains commercial intent, does the person want to buy or sell something?
    - *"I'll buy the Joe-Nastics dvd" (*buy intent)
    - *"I'm selling my black emperor scorpion"* (sell intent)

- Explicit vs. Implicit Intent
  - Does the person explicitly express CI or rather state a possibility in the future?
    - "*Facing Repossession, Let us buy your house for cash now http://tiny.ly/G7Rw*" (Explicitly expresses the intent to buy a house)
    - "*Debating on buying the pair of 80s cop shades..." (*Implicitly states CI as a possibility in the future)

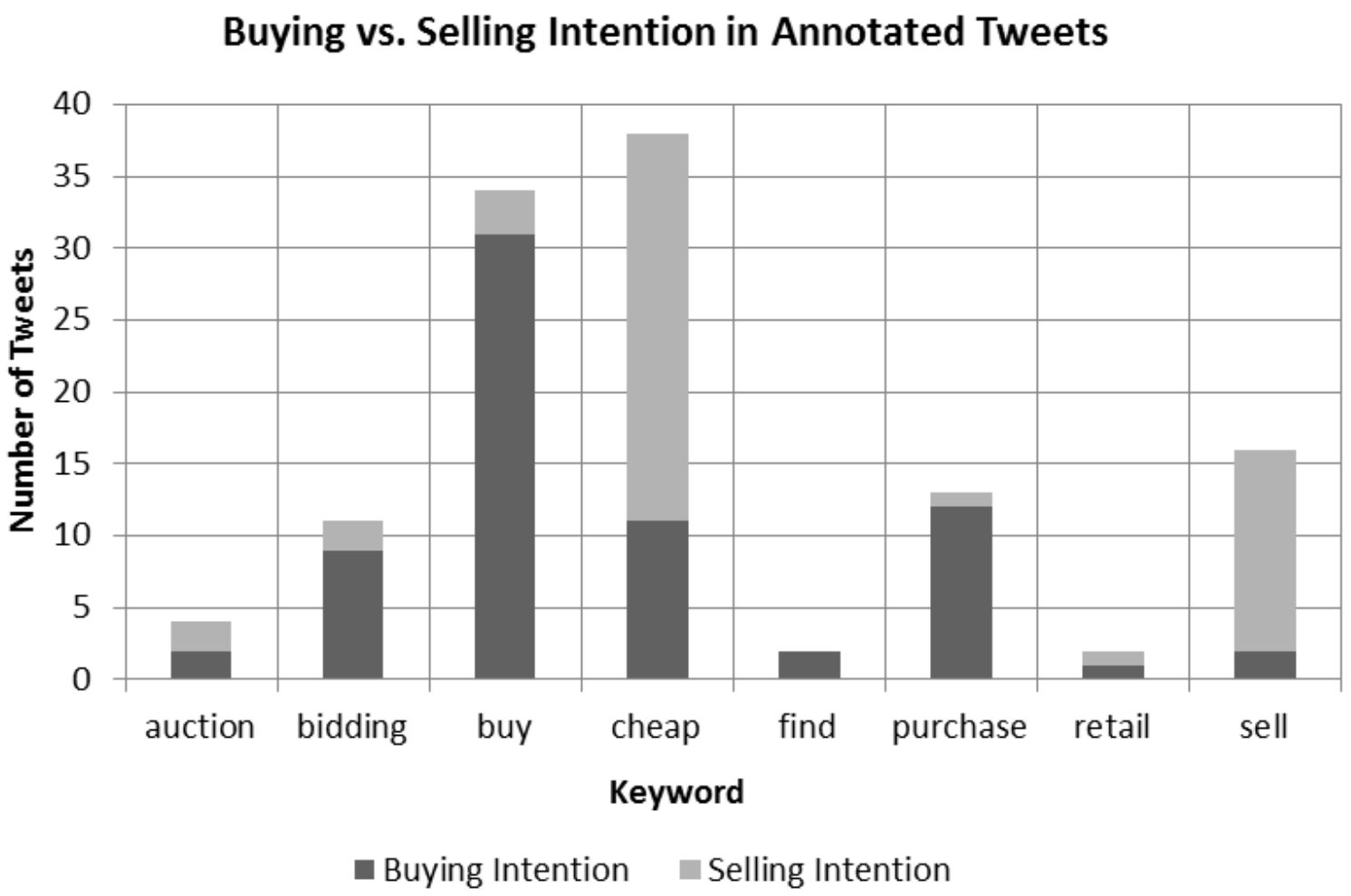
# Collecting and Annotating Tweets

- Generate tweet archives from Twapperkeeper
  - Twapperkeeper is now integrated into HootSuite Archives
- Annotate tweets with regard to commercial intent
  - Keywords: Advertise, Aim to, Auction, Bidding, Buy, Cheap, Cost, Deal, Find, Get, Market, Price, Purchase, Rent, Retail, Sale, Sell
  - 100 tweets each for selected keywords
- Motivation for these keywords
  - Keywords from „Detecting CI in Search Query Logs“ [Dai06]
  - Levin‘s verb classes [Levin93]
  - Randomly chosen tweets had little chance of exhibiting CI

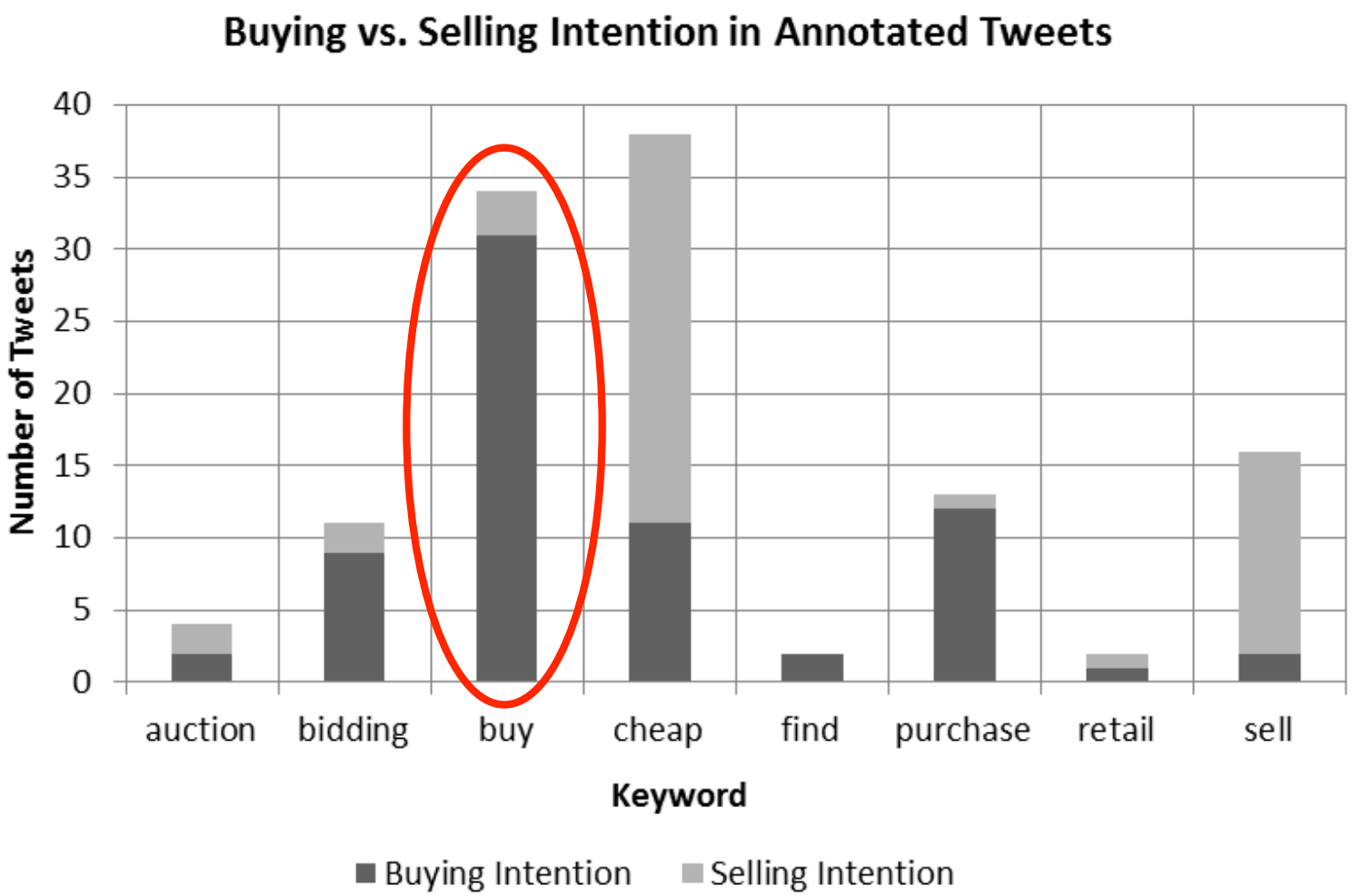
# Percentage of CI per Tweet Archive



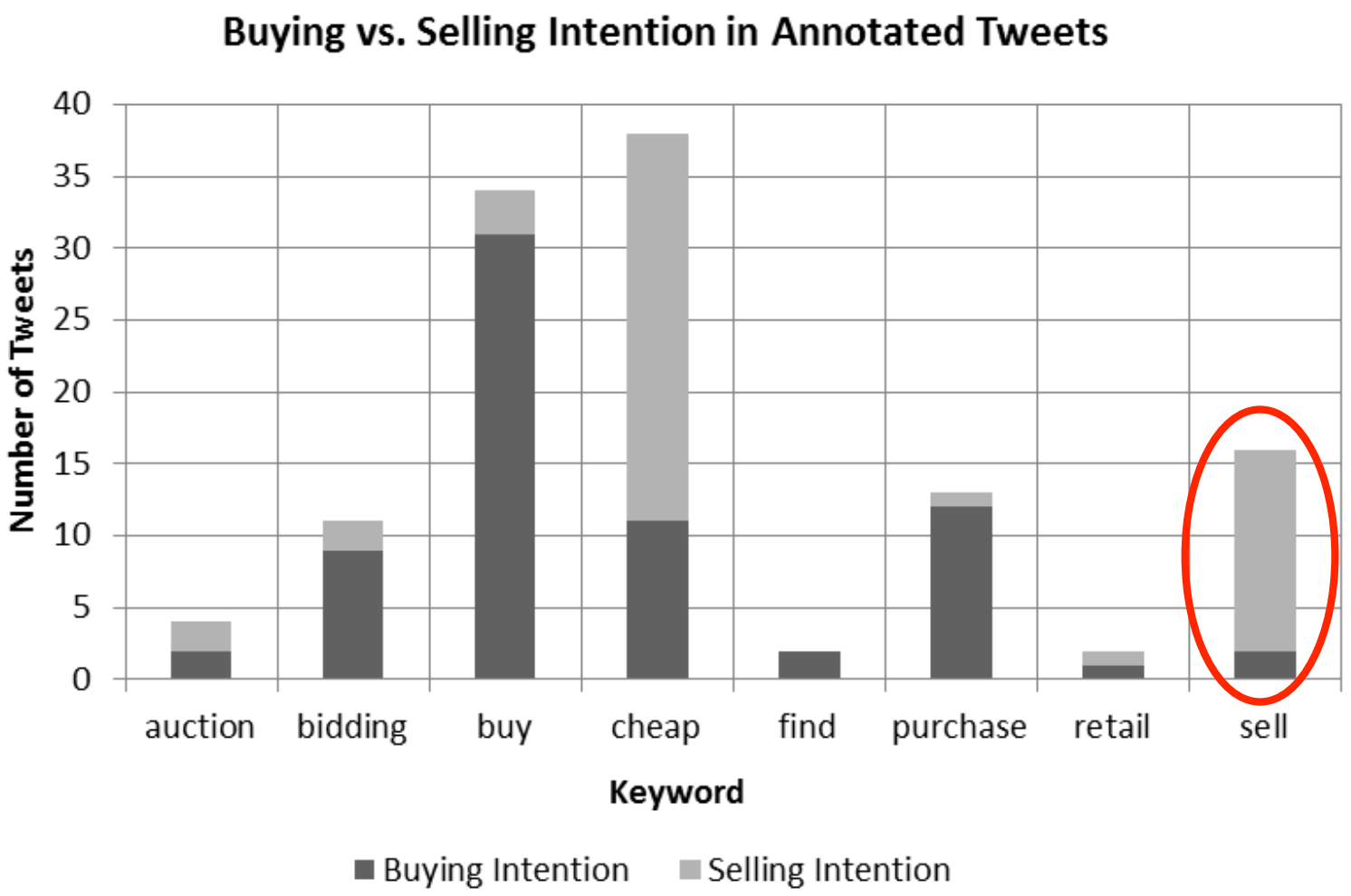
# Buying vs. Selling Intention Percentage



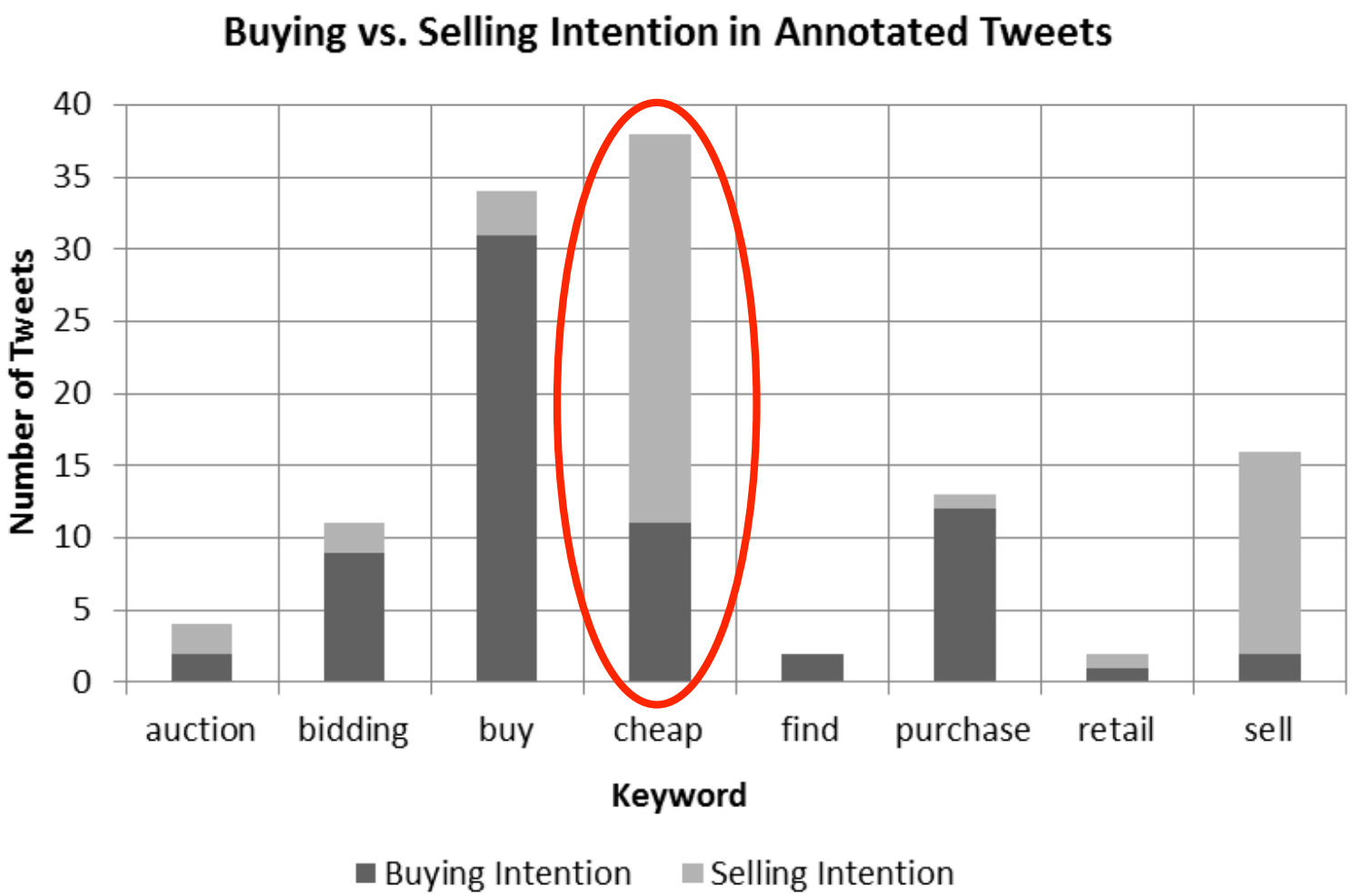
# Buying vs. Selling Intention Percentage



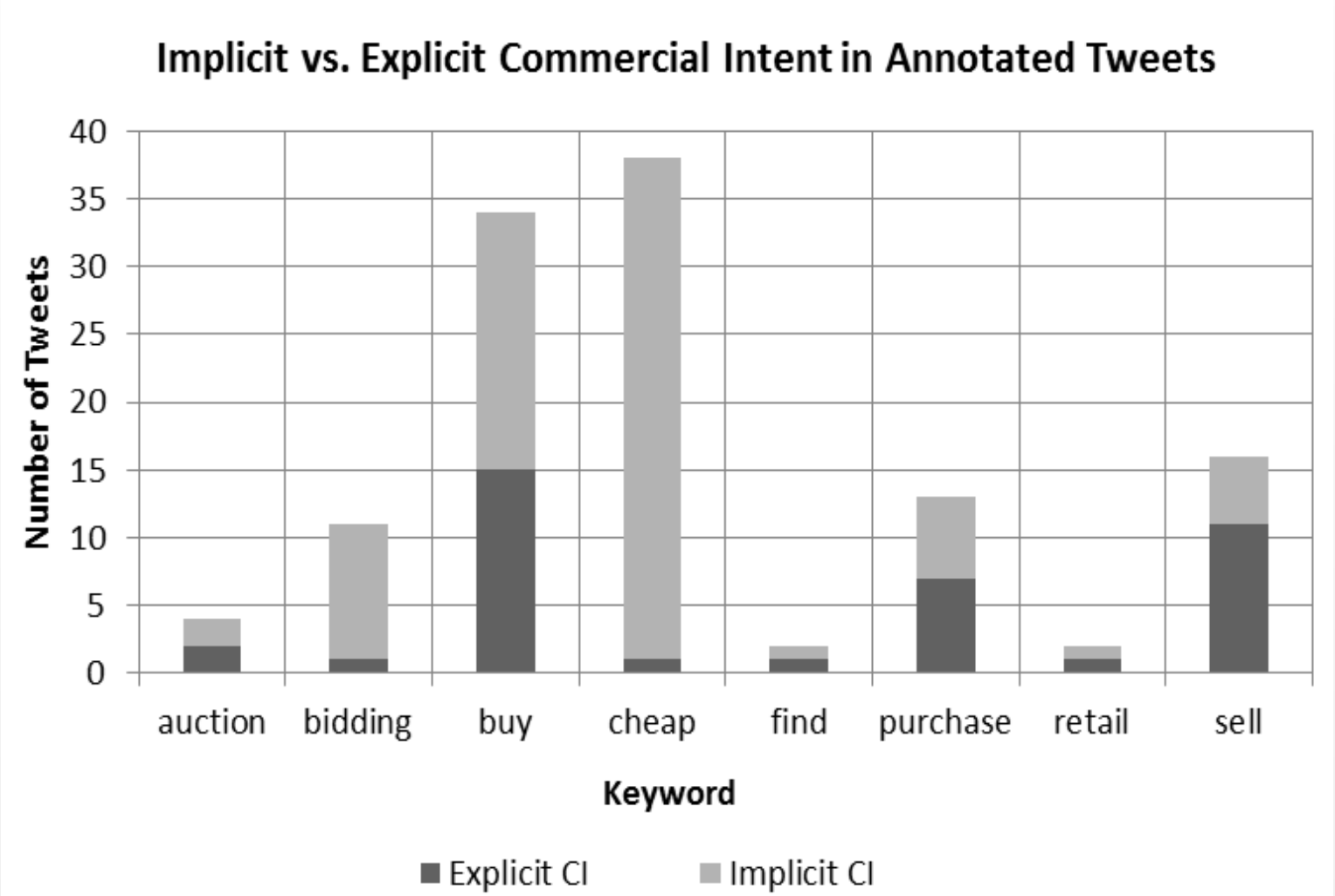
# Buying vs. Selling Intention Percentage



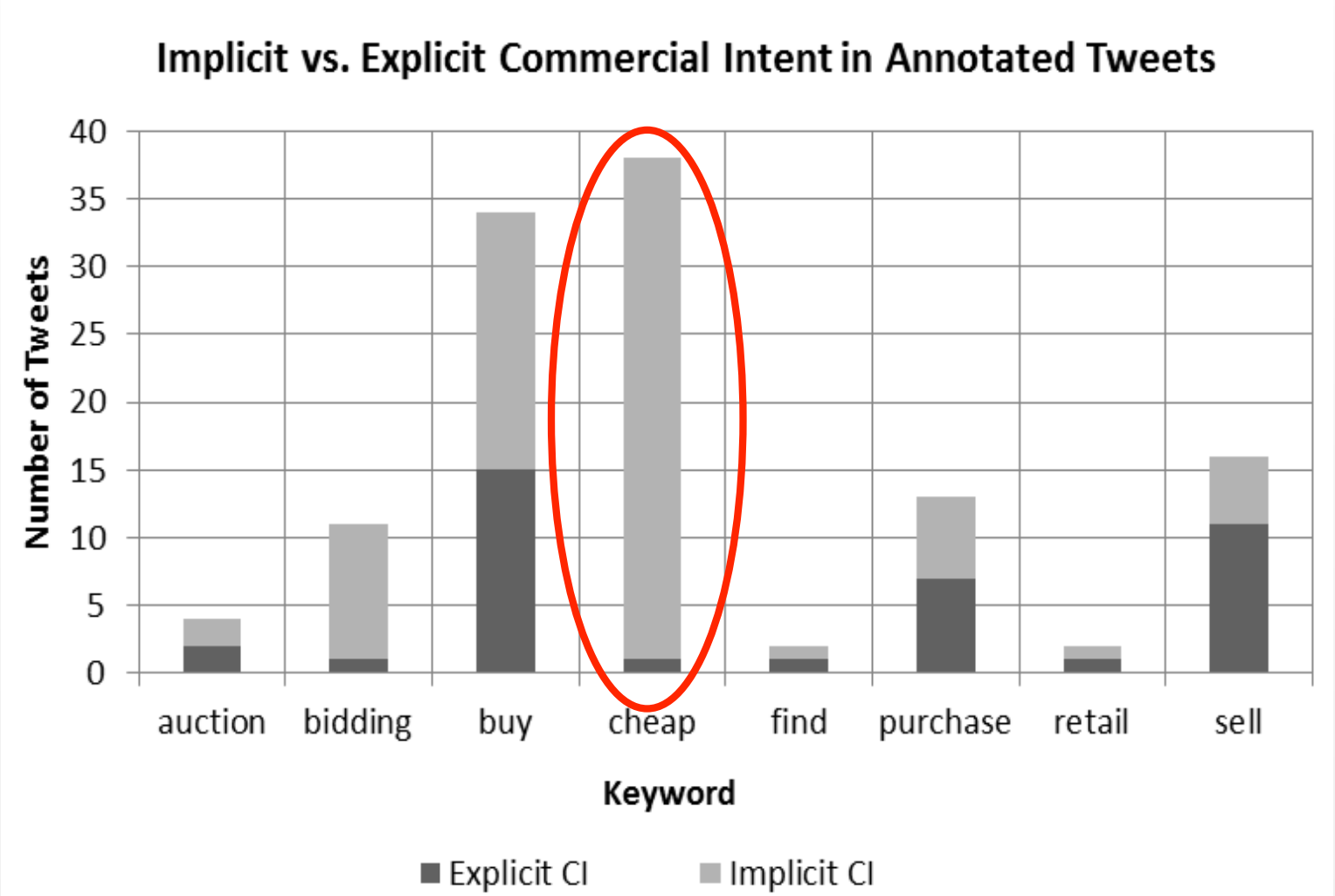
# Buying vs. Selling Intention Percentage



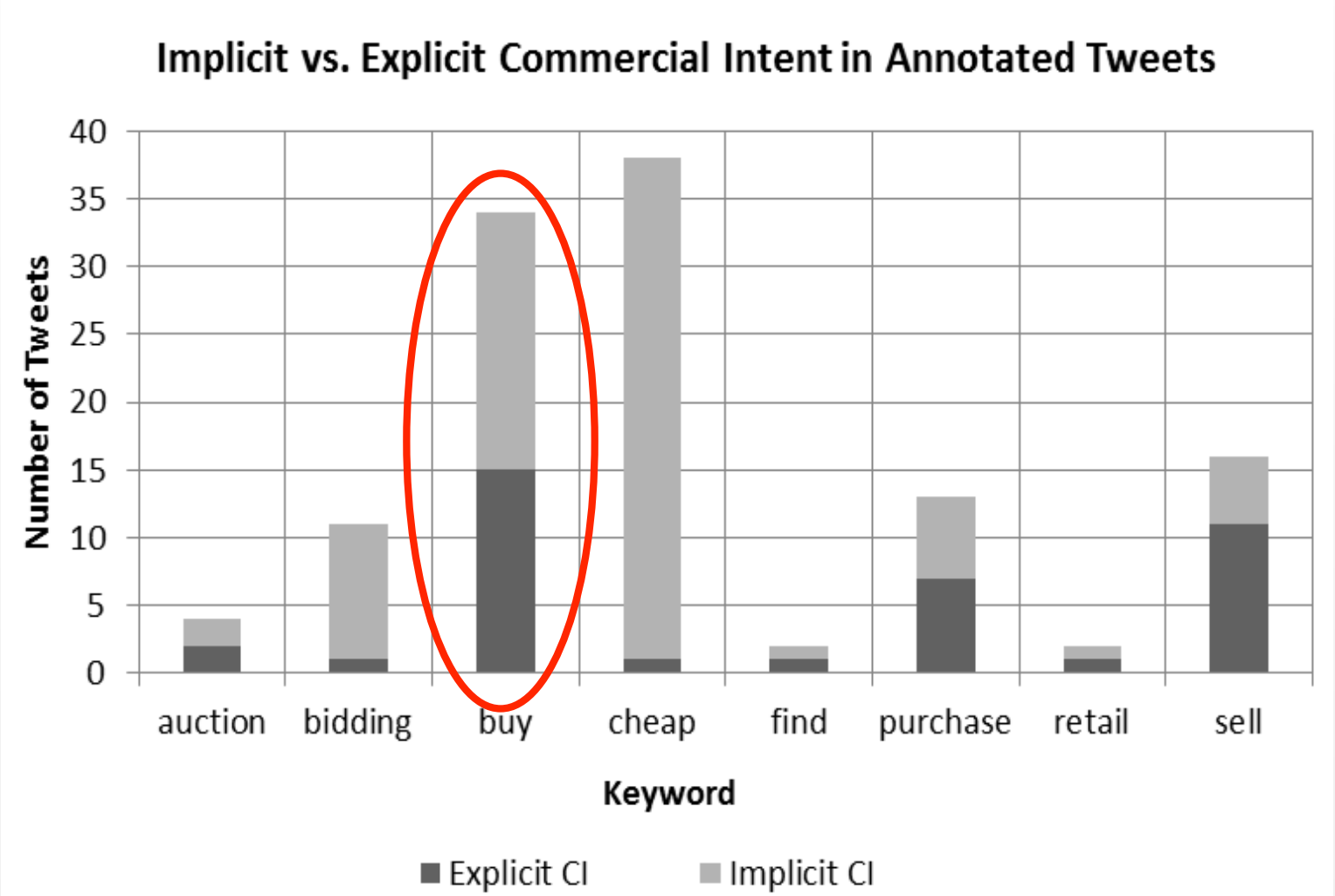
# Explicit vs. Impicit CI Percentage



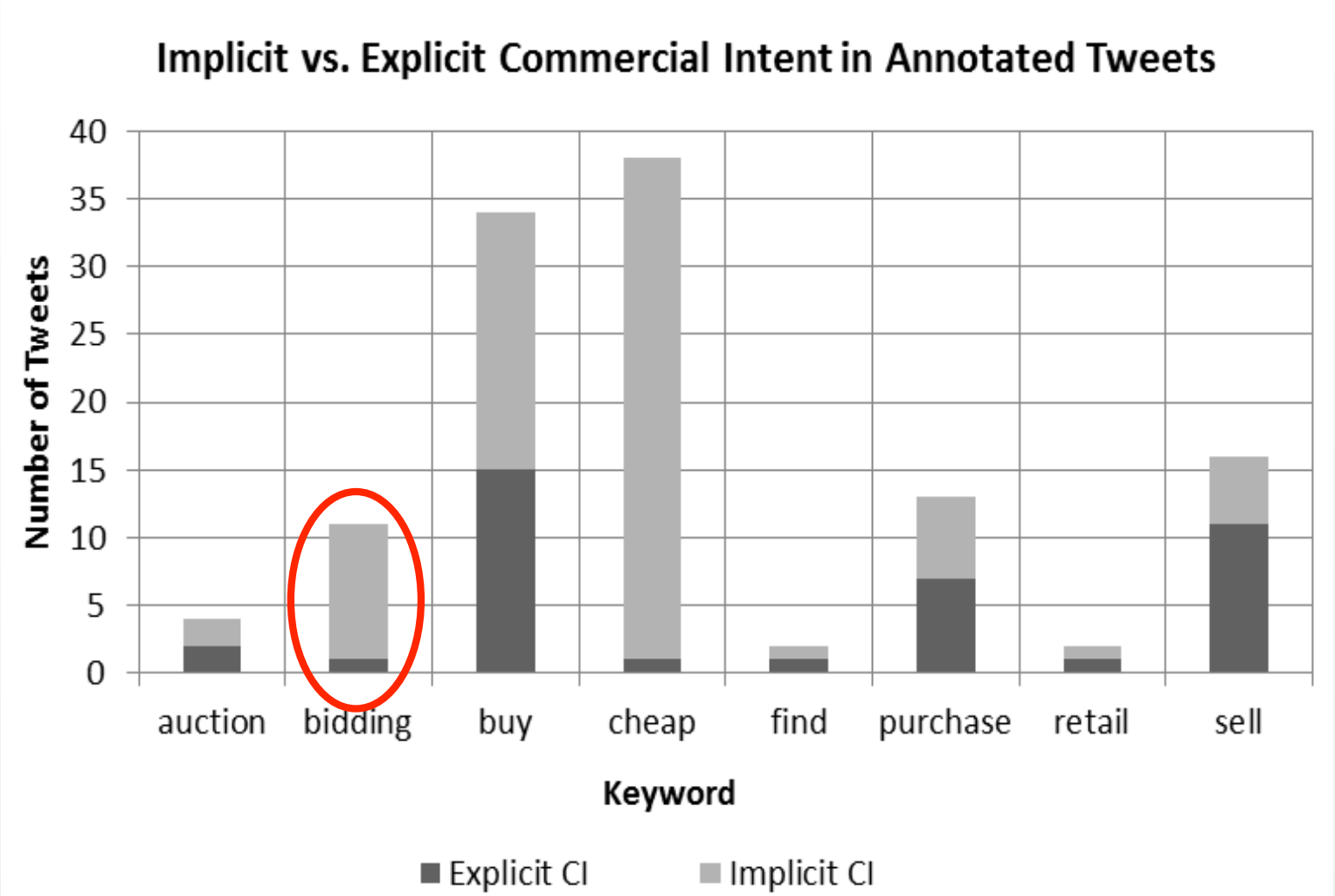
# Explicit vs. Impicit CI Percentage



# Explicit vs. Impicit CI Percentage



# Explicit vs. Impicit CI Percentage



# Experimental Setup

- Preprocess tweets
  - Remove all characters except for A-Z, a-z, 0-9 and spaces
  - Make every character lowercase
  - Replace two or more consecutive spaces with exactly one space
  - Append a period to every tweet
- Remove duplicates
- Apply part-of-speech tagger
- Remove tweets without verb
- Apply WEKA‘s pre-processing suite

# Part-of-Speech Tagger

- Processes sentences and adds suffixes to each word, denoting it as a noun, adjective, preposition, etc.

- E.g. *"Cars_NNS make_VBP Great_JJ Christmas_NNP gifts_NNS !_."*
  - Original: *"Cars make Great Christmas gifts!"*

| Penn Treebank POS Tagset (Excerpt) | | | |
|---|---|---|---|
| **NNP** | proper noun, singular | **JJ** | adjective |
| **NNS** | noun plural | **VBP** | verb, sing. present |

# Feature Generation

- Using WEKA for feature engineering and classification
  - Java-based machine learning toolkit
- With and without stemming
  - SnowballStemmer
- NGramTokenizer (n = 2-5) to generate
  - Textual features
  - Part-of-speech feature

# Top 10 Most Discriminative Attributes

| 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' |
| 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 httpbitlyd3g1e4' |
| 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 amazoncouk httpamznto9hzjhq' |
| 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 httpbitlybzsitp' |

# Training a Classifier

- Training Data
  - Positive class: 158, negative class: 1478
- used WEKA to apply several classification models
  - Linear Support Vector Machines, Nearest Neighbor Algorithm, Decision Trees
- applied 10-fold cross-validation
- Precision, recall and F1-value
  - for positive class

# Classification Model Preliminary Results

| Classifier/Score | Precision | Recall | F1 score |
|---|---|---|---|
| Bayes Complement Naive Bayes | 0,13 | 0,77 | 0,23 |
| BayesianLogisticRegression | 0,57 | 0,08 | 0,14 |
| NaiveBayes | 0,27 | 0,49 | 0,35 |

- Experimented with DMNBtext, J48, NaiveBayesMultinomial, NaiveBayesMultinomialUpdateable, SMO
- POS tags seem not very suitable
  - Use other POS tagger or focus on textual attributes

# Conclusion

- Presented preliminary results for detecting commercial intent on Twitter
  - Next steps to improve the results: examine other feature types: URLs, emoticons, punctuation, uppercase/lowercase, etc.
- Potential applications include
  - Trend detection, i.e. investigate, which products are likely to be bought, which are likely to be sold
  - Improvement of spam detection
  - Additional search facets to Twitter
- Ultimately, our research will help to link buyers and sellers on Twitter and other social networks

# Selected Comments from Reviewers

- Didn‘t you bias your sample?
  - We intentionally did. „Regular“ keywords contain very little to no CI at all and even the top keywords (buy and sell) only contained ~30% CI. As we are interested in the positive class, we aimed to seek out for keywords with a high likelihood of CI.
- Some vagueness in description, e.g. „a suitable amount of tweets“ -> how many?
  - 100 tweets for every archive unless Twapperkeeper didn‘t provide enough. That is 100 tweets for: *advertise, aim to, auction, bidding, buy, cheap, find, purchase, retail, sell.* Fewer for *cost* (21), *deal* (87), *get* (84), *market* (22), *price* (32), *rent* (13) and *sale* (76). Total: 1335 tweets.
- Re-run with additional keywords, e.g. want(ed), swap, trade?
  - We did, this presentation contains the updated results (2000 tweets).

# References


[Ashkan09] Ashkan, A. and Clarke, C. 2009. Term-based commercial intent analysis. In Proc. of the International Conference on Research and Development in Information Retrieval.

[Benczúr07] 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.

[Cohen60] Cohen, J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement.

[Dai06] 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.

[Guo10] 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.

[Kirsh90] Kirsh, D. 1990. When is information explicitly represented? Information, Language and Cognition – The Vancouver Studies in Cognitive Science.

[Kouloumpis11] 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.

[Kröll09] Kröll, M. and Strohmaier, M. 2009. Analyzing human intentions in natural language text. In Proceedings of the International Conference on Knowledge Capture.

[Levin93] Levin, B. 1993. English verb classes and alternations: A preliminary investigation. University of Chicago Press.

[Maynard11] 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.

[Rennie03] 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.

[Sakaki10] 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.

[Strohmaier12] Strohmaier, M. and Kröll, M. 2012. Acquiring knowledge about human goals from search query logs. Information Processing and Management 48, 1.

[Zhao11] 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).

# Thank you for your attention!

# Questions & Answers

Bernd Hollerit
bernd.hollerit@student.tugraz.at
Graz University of Technology, Austria