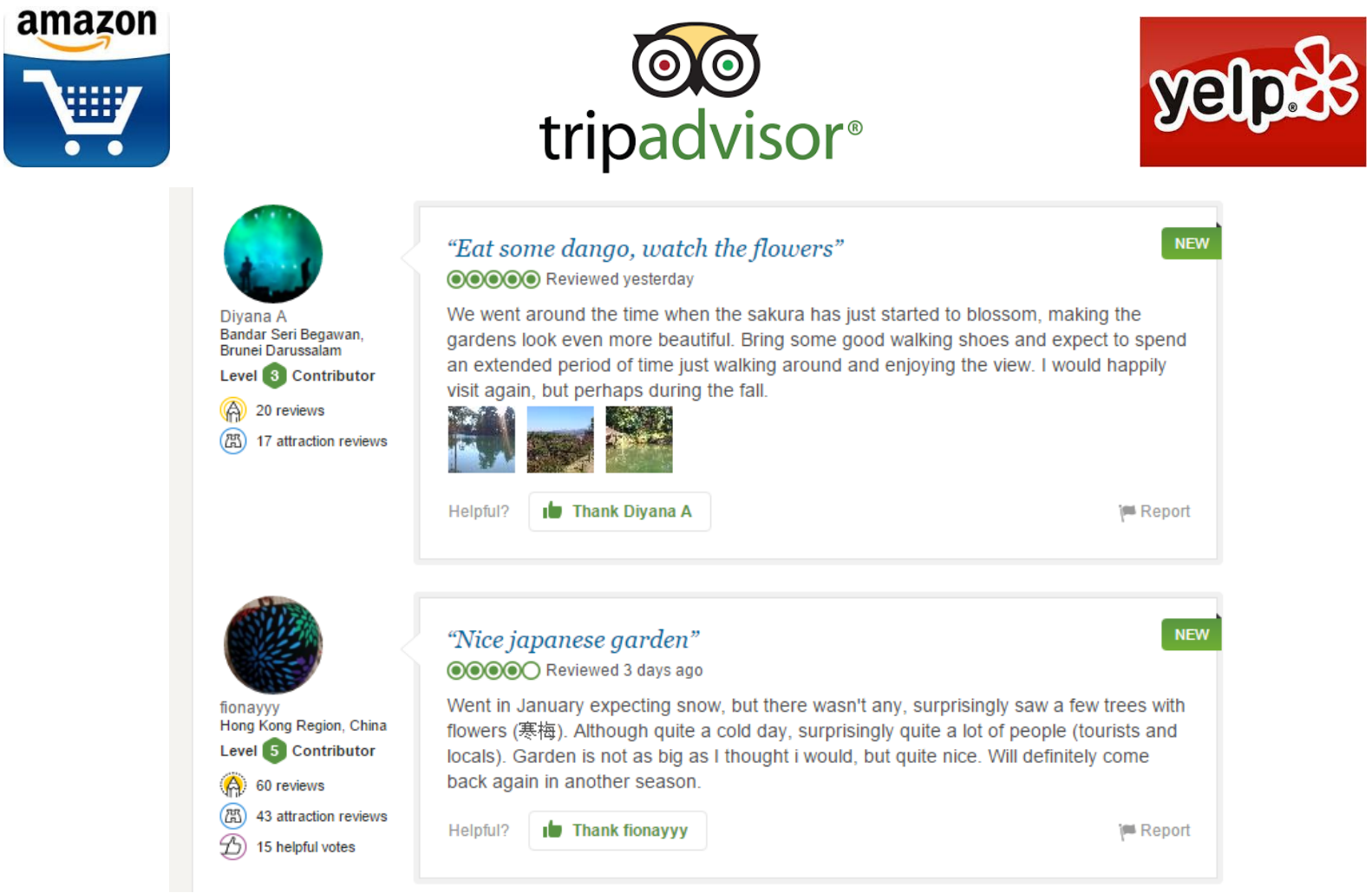
Predicting Review Helpfulness

A Machine Learning&Natural Language Processing based Approach

Ashwin Ittoo
Asst Prof
HEC, Management School,
University of Liège, Belgium

Background

- Online reviews
  - Information source assess product/quality



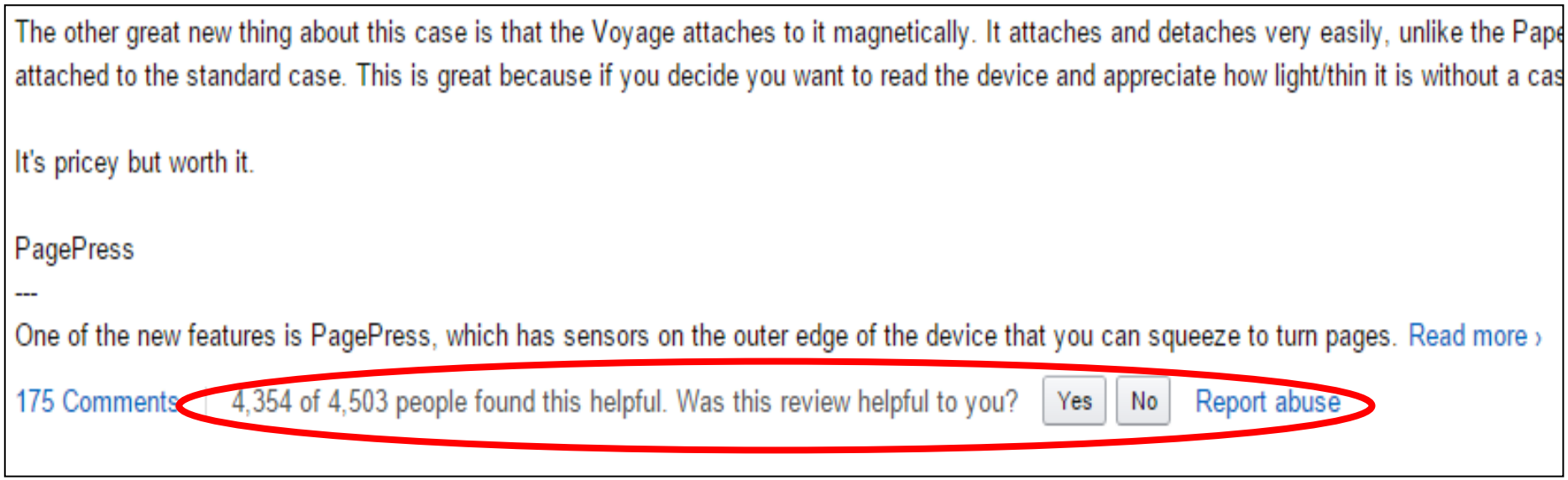
Background (cont)

- However…
- Huge number of human-authored online reviews
  - Hard for consumers to discern really helpful reviews

- Lack of editorial control
  - Great variations in review quality

Background (cont)

- Emergence of online peer-reviewing
  - Consumers vote for review helpfulness



- Helpfulnes votes
  - Indicator of review diagnosticity (Liu & Park, 2015, Mudambi & Schuff, 2010)
  - Facilitate efficient review filtering (Ghose & Ipeirotis, 2008)

Research Questions (RQ)

- « What makes a (customer) review helpful?
  - Fundamental issue yet to be addressed

- Main RQ addressed
  - Linguistic devices contributing to review helpfulness
  - Predicting reviewing helpfulness

- Approach based on
  - Linguistics, Natural Language Processing (NLP)
  - Machine Learning (ML)

- Development of novel algorithms
  - Automatic detection of linguistic structrures
  - Predict review helpfulness

Motivations

- Automatically assessing review helpfulness
- Practical/Industrial motivations
  - Positively impact customer purchase decisions
  - Improve website social presence; attract customers, increase sales (Kumar & Benbasat 2006)
- Scientific motivations
  - Novel NLP/ML algorithms
  - Assessing deeply embedded textual features (helpfulness)

Presentation Structure

- Related studies
  - Automatically assessing review quality
- Proposed approach
  - Feature engineering
  - Machine learning experiments
  - Results
  - On-going/future work

Related Studies – Review Helpfulness

- 3 main areas of related resarch

1. Spam, fake review detection
   - Detecting duplicate reviews (Jindal&Liu, 2007)
   - Pattern-based spam detection (Mukherjee et al., 2013)

2. Influence of reviews on sales
   - Positive reviews and book sales positively correlated (Chevalier & Mayzlin, 2006)
   - Positive reviews, average review ratings and box-office sales positively correlated (Dellorcas et al., 2007)

Related Studies – Review Helpfulness

- Clarification of extant literature
- 3 main areas of related resarch

1. Spam, fake review detection
   - Detecting duplicate reviews (Jindal&Liu, 2007)
   - Pattern-based spam detection (Mukherjee et al., 2013)

2. Influence of reviews on sales
   - Positive reviews and book sales positively correlated (Chevalier & Mayzlin, 2006)
   - Positive reviews, average review ratings and box-office sales positively correlated (Dellorcas et al., 2007)

Related Studies – Review Helpfulness (cont)

3. Predicting review helpfulness/quality

- Factors contributing to review helpfulness
- Identity disclosure (Sussman et al., 2003; Forman et al., 2008)
  - Improves review credibility; more helpful review
  - Presence of photos, email addresses, …
- Reviewers' characteristics (Gilly et al., 1998, Liu&Park, 2015)
  - Expertise, reputation
  - Review from expert likelier to influence purchase decision, more helpful
  - Number of followers/friends, number of reviews, replies
- Review star ratings (Danescu et al., 2007)
  - Reviews with extreme ratings (1,4) more helpful
  - Various statistics (mean, variance) from star ratings
- Perceived enjoyment in reading review (Liu&Park, 2015)
  - Number of clicks « Cool » button (only for Yelp! Data)

Sort by: **Date**

Courtesy Plumbers
Contractors, Plumbing
601 E Edna Pl
Covina, CA 91723

4/18/2013

The work was done in a timely manner. A call back was necessary - the follow up was above average - both customer service people were very courtesy.The workman Tom call back within a reasonable time. They are quarantee their work. We can recommend them and will use them again.

Was this review ...?

Useful Funny Cool

Related Studies – Review Helpfulness (cont)

- Shortcomings of existing research

- Extrinsic factors (identity disclosure, reviewers' characteristics)
  - Not accurately predictive of review helpfulness

- Factors hard to operationalize, formalize, measure
  - Coarse-grained approximates, oversimplify actual values
  - Identity disclosure as binary {0,1}
  - Expertise as number of friends

- Factors not always available
  - Number of fans, « cool » button

Linguistic factors for Review Helpfulness

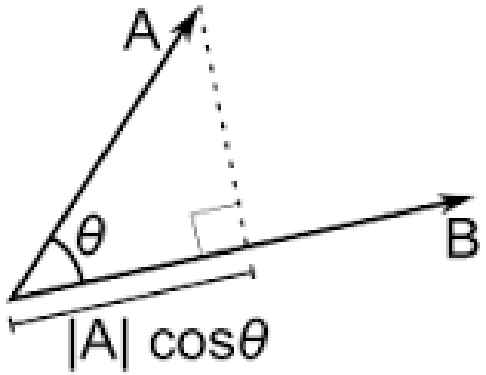
- Reviews' textual contents, styles
  - Intrinsic factor, inherent to reviews

- Theoretical underpinnings
  - Textual contents, styles determine message persuasiveness (Schindler, & Bickart, 2012)
  - "Source effect" of message (Janis et al. 1959)

Linguistic factors for Review Helpfulness

- Zhang & Varadarajan (2006) predicting Amazon review helpfulness
  - Cosine similarity schemes on TF IDF scores
  - Shallow syntactic features (surface forms of words, lemmas, parts of speech tags)

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$



$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Linguistic factors for Review Helpfulness (cont)

- Review elaborateness to predict helpfulness (Mudambi&Schuff, 2010)
  - More elaborate → more helpful
  - Elaborateness = number of words
- Stylometric features (Viswanathan et al., Lee&Choeh, 2014)
  - N-grams, syntactic dependencies, sentiments
- However
  - Lack scientific basis
  - Ad-hoc features
  - No consensus in feature definition

Proposed Approach

- Grounded in ***argumentation theory***

- Hypothesis
  - Argumentative structures useful predictor of review helpfulness

- Rationale for Argument as predictor:
  - ***Pragma-Dialectical theory*** (Sycara, 1990 , Van Eemeren, & Houtlosser,2003)
  - Argument as linguistic device for persuasion
  - Found in dialogue, where one party convinces another
  - Reader votes review as helpful only if persuaded of utility of review

What is an “Argument”?

- Argument (Palau & Moens, 2009)
  - Set of premises/evidence/facts
  - Claim/conclusion
  - Claim: proposition, either true or false

- 2 steps in proposed approach
  1. Identify arguments from review texts
  2. Determine whether arguments effectively predict review helpfulness

Argumentation Theories

- Several argumentation theories

- Toulmi's formalization (Toulmi, 2003)
  - 6 components of arguments: claim, data, warrant, backing, qualifier, rebuttal
  - Complex formalization, difficult to "learn" and to identify in text (ML, NLP)

- Rhetorical Structure Theory (RST)
  - Several rhetorical relationships between 2 text spans; nucleus-satellite
  - Nucleus: central to the discourse
  - Satellite: interpretable only w.r.t. nucleus
  - ***More suitable for computational purposes***

Formalizing the "Argument"

- 2 components of argument
  1. Premise
  2. Conclusion

- Components essential in human cognition (Paulus & Moens, 2009)
  - Distinguish argumentative structures vs. normal statements

- Premise-conclusion relation
  - Various syntactic mechanisms
  - Subordination, coordination

- [I love this book]$_{\textit{conclusion,}}$ as [it reflects the authors' trials and tribulations....] $_{\textit{premise}}$

Feature Engineering

- Formal definition for defining lexical features
  - Signal argumentative structures in text

- Prominent feature from discourse theory
  - Discourse connectives
  - [I love this book]$_{\textit{conclusion}}$, ***as*** [it reflects the authors' trials and tribulat1ions….] $_{\textit{premise}}$

Feature Engineering (cont)

- Subordinating conjunctions (because, since,…)
  - ***Since*** I am a fan of autobiography, I could not resist but to buy …
- Coordinating conjunctions (but, and, or,…)
  - *This is not a best-seller, ***but*** I still love…
- Adverbials (however, as a result, otherwise…)
  - I had a bad experience with Motorola. ***As a result***, I switched to Samsung…
- Conjoined connectives (when-if, if-when)
  - I will be satisfied ***if and when*** I receive a refund…

Experiments

- 3 datasets of consumer reviews
  1. Amazon.com (books, movies)
  2. Yelp! (restaurants)
  3. TripAdvisor (restaurant, hotels)

- Standard, used in previous research (Viswanathan et al., Jindal&Liu, 2008, Wang et al., 2011)

- 100,000 reviews randomly selected from each dataset

Experiments – Data Cleaning

- Data in JSON/XML format

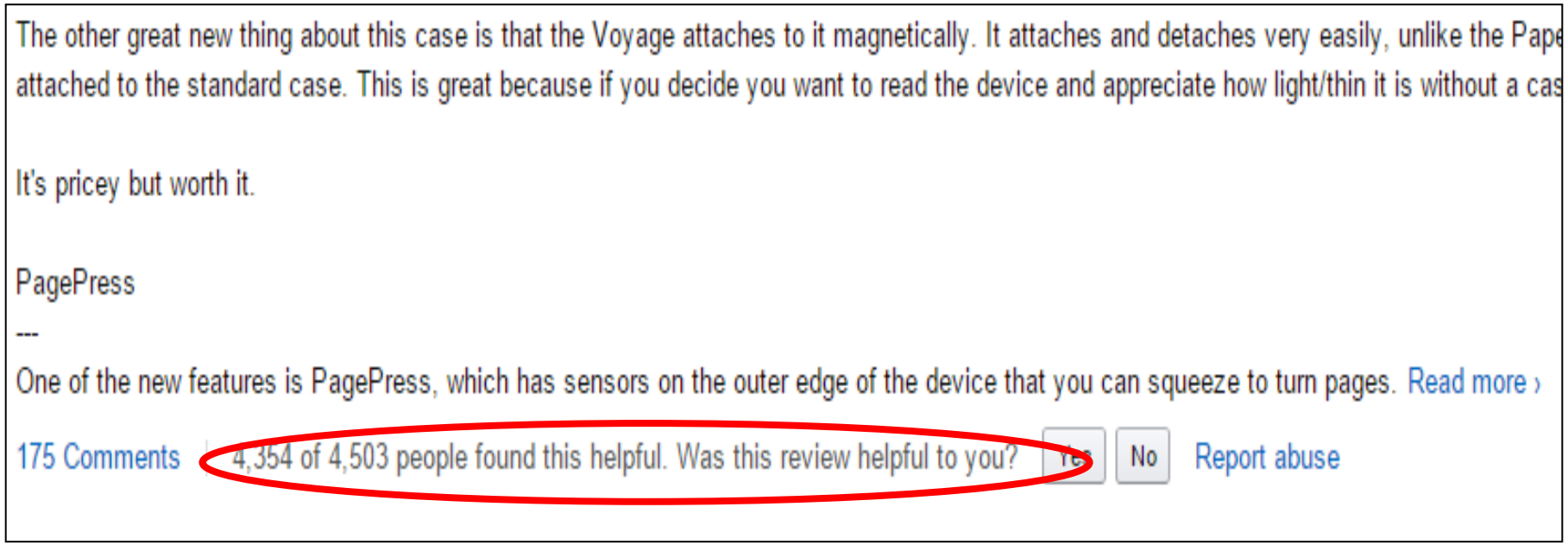
review

```
{
    'type': 'review',
    'business_id': (encrypted business id),
    'user_id': (encrypted user id),
    'stars': (star rating, rounded to half-stars),
    'text': (review text),
    'date': (date, formatted like '2012-03-14'),
    'votes': {(vote type): (count)},
}
```

- Meta-data of interest
  - Review text
  - Number of votes (review helpfulness)

Experiments – Data Cleaning (cont)

- Following previous studies (Ghose, A., & Ipeirotis , 2011)
- If >=60% of users found review X helpful
  - Then, classify X as “true” (helpful)
- Else
  - Classify X as “false” (not helpful)



Experiments – Data Cleaning (cont)

- Text data inherently noisy
- Requires further cleaning/pre-processing
  - Stopwords removal
  - Morphological analysis (lemmatization), e.g. “loved”, “love”, “loves” → “love”

- Reduce dimensionality via feature selection
- Best results with cube mutual information

$$\log \frac{\left(\frac{f(x,y)}{N}\right)^3}{\frac{f(x)}{N} \times \frac{f(y)}{N}}$$

- Degree of correlation between word x (e.g. “book”) in the review and class y (helpful/not helpful) of the review
- Result: find best words characterizing each class

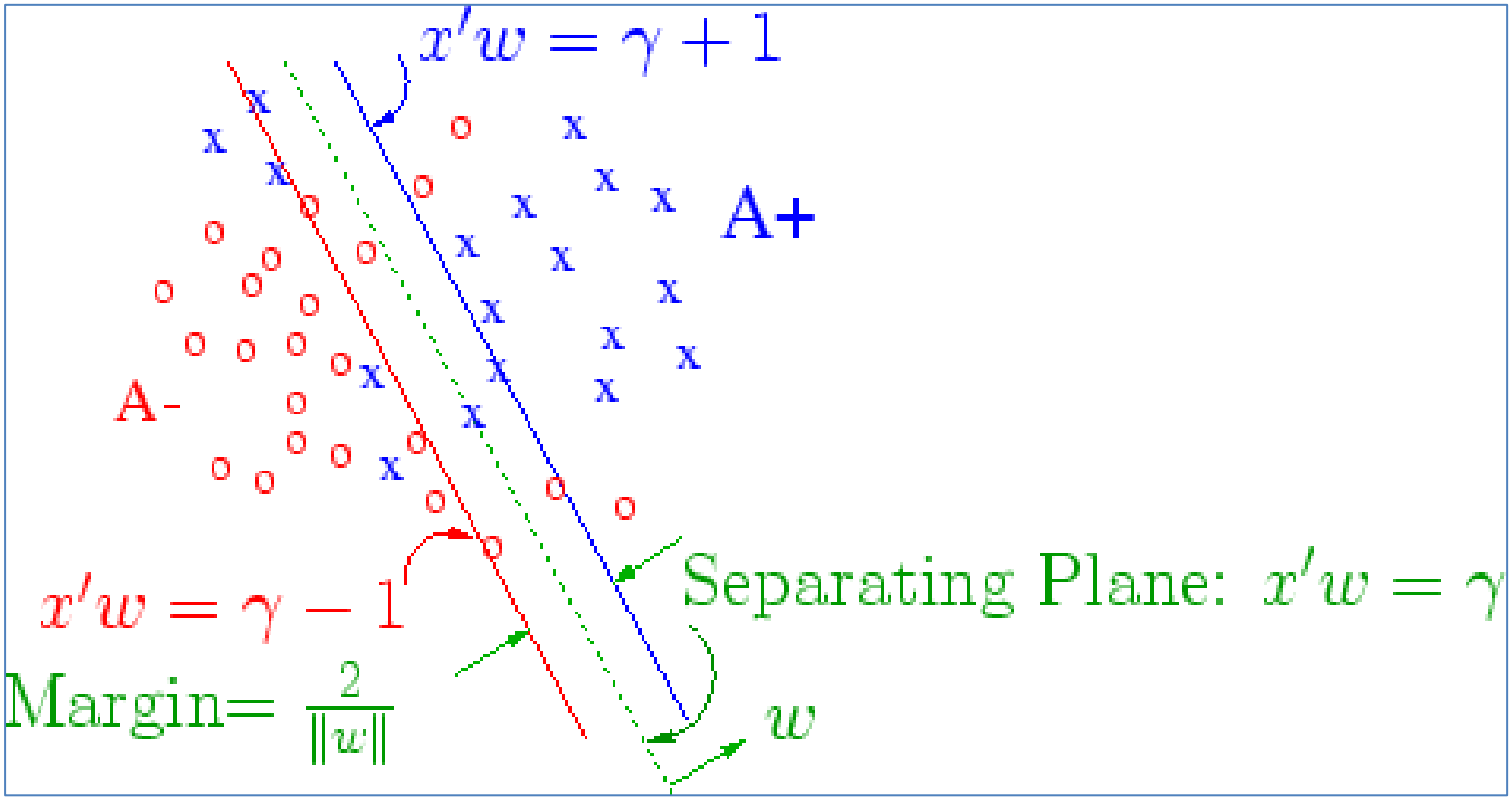
Experiments – Machine Learning

- Adopted ML formulation
  - Given a review,
  - Predict its likeliest class (helpful or not)

- Various ML paradigms/classifiers investigated
  - ZeroR
  - SVM
  - Random Forest (RF)

Experiments – Machine Learning (cont)

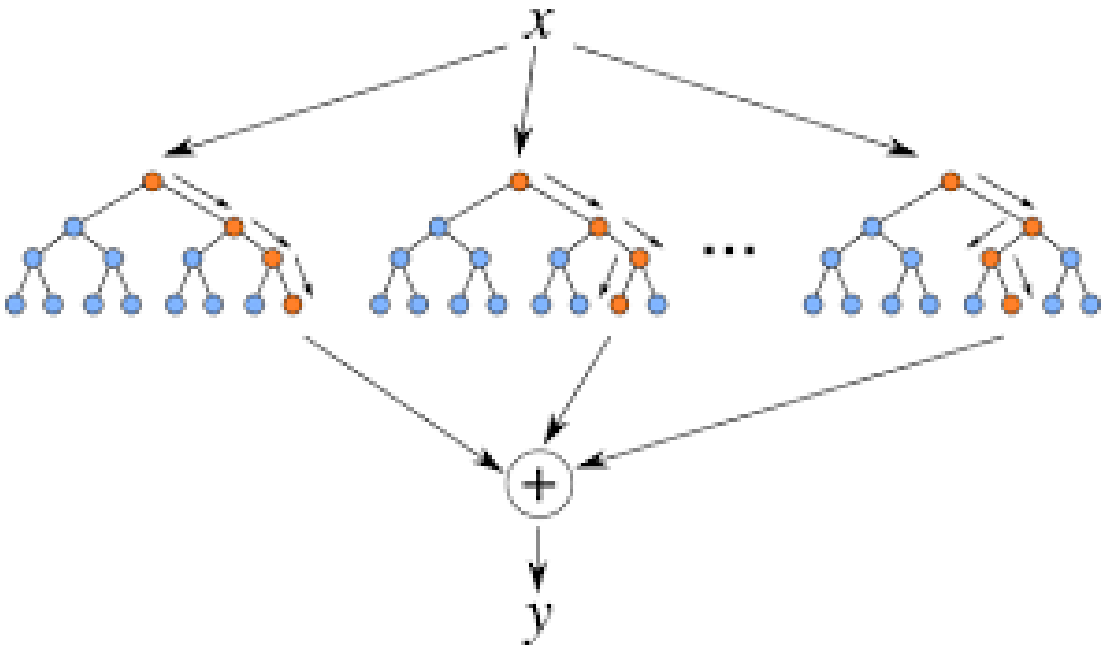
- SVM classifier
  - Optimal hyperplane separating 2 classes



x:helpful
o:not helpful

Experiments – Machine Learning (cont)

- RF classifier
  - Combines predictions of multiple decision tree
  - Makes best possible prediction



Experiments - Class Imbalance Issue

- Datasets highly imbalanced
  - Helpful class >> not helpful class
- Hard predicting minority class (not helpful)
- High accuracy with dumb classifier
  - If 80% of data belongs to helpful class (20% not helpful)
  - Accuracy(Predicting everything as helpful)=0.8
  - Results are misleading
- Need to generate a balanced distribution

Experiments - Class Imbalance Issue (cont)

- SMOTE (Synthetic Minority Oversampling Technique) (Chawla et al. 2002)
  - Resolve class imbalance issue
  - Generates synthetic (artificial) examples of minority samples
  - From K-Nearest Neighbors

- Operates in feature space
- Better performance vs. under/oversampling techniques in data space
  - E.g. replicate minority class examples

Experiments – Set up

- Baseline
  - For comparing contribution of argumentative patterns
  - Helpfulness prediction with argumentative patterns vs. with baseline features

- Each classifier (SVM, RF) evaluated over data twice
  - Setup1: using baseline features
  - Setup2: using argumentative patterns
- Measure accuracy of helpfulness prediction (per setup)
- Determine contribution of argumentative patterns

Experiments – Baseline Features

- Features employed in previous research
- Star rating (numerical, 1-5)
- Extreme rating
  - 1 iff rating $\in \{1,5\}$, 0 otherwise
- Avg. sentence length
- Number of 1st, 2nd pronouns
- Sentiment scores (numerical, computed from SentiWordNet)

Experiments – Baseline Features (cont)

- Baseline features: Readability metrics
- SMOG (Mc Laughlin, 1969)
  - Polysyllables: words >= 3 syllables

$$1.0430\sqrt{\text{number of polysyllables} \times \frac{30}{\text{number of sentences}}} + 3.1291$$

- Flesch-Kincaid (Kincaid et al., 1975)

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

- Gunning fog (Gunning, 1952)
  - Complex words = polysyllables

$$0.4\left[\left(\frac{\text{words}}{\text{sentences}}\right) + 100\left(\frac{\text{complex words}}{\text{words}}\right)\right]$$

- Coleman-Liau (Coleman&Liau, 1975)
  - L = avg. number of letters per 100 words
  - S = avg. number of sentences per 100 words

$$0.0588L - 0.296S - 15.8$$

Measuring Prediction Accuracy

- Standard metrics
  - Precision(P)

$$\frac{true\_positive}{true\_positive + false\_positive}$$

  - Recall (R)

$$\frac{true\_positive}{true\_positive + false\_negative}$$

  - F1-score

$$\frac{2 \times P \times R}{P + R}$$

- Estimated via 10-fold cross-validation
  - Divide data into 10 subsamples
  - Use 9/10 for training
  - Test over 1/10 (no SMOTE on test set)

Experiments - Results

- Best results with RF
  - Outperforms SVM in predicting review helpful/not
  - Over all 3 datasets

Amazon

Feature set	Precision (true)	Recall (true)	F1 (true)	Precision (false)	Recall (false)	F1 (false)	Precision (overall)	Recall (overall)	F1 (overall)
Rating + length + readability	0.734	0.77	0.752	0.758	0.721	0.739	0.746	0.746	0.745
Rating + length + extremity + readability	0.758	0.727	0.742	0.758	0.727	0.742	0.748	0.747	0.747
Baseline	0.736	0.779	0.757	0.765	0.721	0.743	0.751	0.75	0.75
Argumentation	0.78	0.78	0.78	0.77	0.76	0.765	0.775	0.77	0.77
Baseline + argumentation	0.81	0.79	0.80	0.78	0.76	0.77	0.795	0.775	0.78

Tripadvisor

Feature set	Precision (true)	Recall (true)	F1 (true)	Precision (false)	Recall (false)	F1 (false)	Precision (overall)	Recall (overall)	F1 (overall)
Rating + length + readability	0.733	0.772	0.752	0.759	0.719	0.738	0.746	0.745	0.745
Rating + length + extremity + readability	0.73	0.778	0.753	0.763	0.713	0.737	0.746	0.745	0.745
Baseline	0.735	0.788	0.761	0.771	0.717	0.743	0.753	0.752	0.752
Argumentation	0.765	0.78	0.77	0.78	0.73	0.75	0.77	0.76	0.76
Baseline + argumentation	0.77	0.80	0.78	0.76	0.75	0.75	0.765	0.775	0.77

Yelp

Feature set	Precision (true)	Recall (true)	F1 (true)	Precision (false)	Recall (false)	F1 (false)	Precision (overall)	Recall (overall)	F1 (overall)
Rating + length + readability	0.686	0.738	0.711	0.716	0.662	0.688	0.701	0.7	0.7
Rating + length + extremity + readability	0.692	0.743	0.716	0.722	0.669	0.694	0.707	0.706	0.705
Baseline	0.711	0.754	0.732	0.738	0.693	0.715	0.725	0.724	0.724
Argumentation	0.745	0.76	0.75	0.75	0.736	0.75	0.75	0.748	0.75
Baseline + argumentation	0.75	0.77	0.76	0.75	0.75	0.75	0.75	0.76	0.75

Results' Discussion

- $F1_{argumentation\_features} > F1_{baseline\_features}$
- $F1_{argumentation\_features} \sim F1_{baseline\_features+argumentation\ patterns}$

- ***Argumentation patterns useful predictor of review helpfulness***

Ongoing/Future Work

- Argumentation Mining algorithms
  - Automatically detecting argumentation structures (premises, conclusions) from review texts
  - Classifying premises (support or attack)
- Incorporate meta-data as features
  - Product type (hedonistic vs. utilitarian), age of reviews,…
- Deep Learning with Neural Networks
  - Move from word embedding to argument embedding

ありがとう

References

- Berger, C. R., & Calabrese, R. J. (1975). Some explorations in initial interaction and beyond: Toward a developmental theory of interpersonal communication. *Human communication research*, *1*(2), 99-112.
- Liu, Z., & Park, S. (2015). What makes a useful online review? Implication for travel product websites. *Tourism Management*, *47*, 140-151.
- Mudambi, S. M., & Schuff, D. (2010). What makes a helpful review? A study of customer reviews on Amazon. com. *MIS quarterly*, *34*(1), 185-200.
- Ghose, A., & Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *Knowledge and Data Engineering, IEEE Transactions on*,*23*(10), 1498-1512.
- Kumar, N., & Benbasat, I. (2006). Research note: the influence of recommendations and consumer reviews on evaluations of websites.*Information Systems Research*, *17*(4), 425-439.
- Jindal, N., & Liu, B. (2007, May). Review spam detection. In *Proceedings of the 16th international conference on World Wide Web* (pp. 1189-1190). ACM.
- Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. S. (2013, July). What yelp fake review filter might be doing?. In *ICWSM*.
- Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of marketing research*, *43*(3), 345-354.

References (cont)

- Dellarocas, C., Zhang, X. M., & Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures.*Journal of Interactive marketing*, *21*(4), 23-45.
- Sussman, S. W., & Siegal, W. S. (2003). Informational influence in organizations: An integrated approach to knowledge adoption. *Information systems research*, *14*(1), 47-65.
- Forman, C., Ghose, A., & Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research*, *19*(3), 291-313.
- Gilly, M. C., Graham, J. L., Wolfinbarger, M. F., & Yale, L. J. (1998). A dyadic study of interpersonal information search. *Journal of the Academy of Marketing Science*, *26*(2), 83-100.
- Danescu-Niculescu-Mizil, C., Kossinets, G., Kleinberg, J., & Lee, L. (2009, April). How opinions are received by online communities: a case study on amazon. com helpfulness votes. In *Proceedings of the 18th international conference on World wide web* (pp. 141-150). ACM.
- Janis, I.L.,, Hovland, C. I., Field, P. B., Linton, H., Graham, E., Cohen, A. R., Rife, D., ... & King, B. T. (1959). *Personality and persuasibility* (pp. 281-299). New Haven: Yale University Press.
- Schindler, R. M., & Bickart, B. (2012). Perceived helpfulness of online consumer reviews: the role of message content and style. *Journal of Consumer Behaviour*, *11*(3), 234-243.
- Zhang, Z., & Varadarajan, B. (2006, November). Utility scoring of product reviews. In *Proceedings of the 15th ACM international conference on Information and knowledge management* (pp. 51-57). ACM.

References (cont)

- Viswanathan, V., Mooney, R., & Ghosh, J. Detecting Useful Business Reviews using Stylometric Features.
- Lee, S., & Choeh, J. Y. (2014). Predicting the helpfulness of online reviews using multilayer perceptron neural networks. *Expert Systems with Applications*, *41*(6), 3041-3046.
- Sycara, K. P. (1990). Persuasive argumentation in negotiation. *Theory and decision*, *28*(3), 203-242.
- Palau, R. M., & Moens, M. F. (2009, June). Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law* (pp. 98-107). ACM.
- Toulmin, S. E. (2003). *The uses of argument*. Cambridge University Press.
- Van Eemeren, F. H., & Houtlosser, P. (2003). The development of the pragma-dialectical approach to argumentation. *Argumentation*, *17*(4), 387-403.
- Ma, M., Huang, L., Xiang, B., & Zhou, B. (2015). Dependency-based Convolutional Neural Networks for Sentence Embedding. *Volume 2: Short Papers*, 174.

References (cont)

- Hongning Wang, Yue Lu and ChengXiang Zhai. Latent Aspect Rating Analysis without Aspect Keyword Supervision. The 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'2011), P618-626, 2011
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 321-357.
- Mc Laughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of reading*, *12*(8), 639-646.
- Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975).*Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel* (No. RBR-8-75). Naval Technical Training Command Millington TN Research Branch.
- Gunning, Robert (1952). The Technique of Clear Writing. McGraw-Hill. pp. 36–37.
- Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, *60*(2), 283.