Words Can Weight

Aaron F. McKenny
August 9, 2019
What would the response be...?

-I love you...
-I like you...
Words have both meaning and weight

- “I love you”
- “Dammit Janet”
- “We are an innovative company”
- “I like you”
- “Gosh Janet”
- “We are a software company”
Review of management CATA research

- Journals
  - 12 usual suspects (AMJ, ASQ, JAP…)

- Years
  - 2000-2018

- Search criteria (18)
  - Technique ("CATA", “computerized text”, “computer-aided text”…)
  - Tools ("LIWC", “Diction”, “CAT Scanner”…)
  - Process (“dictionary”, “word list”, “word count”…)

Review of management CATA research

- Initial sample: 167
- Use dictionary-based coding: 124 (74%)

- Report that weights were used… 4 (2%)
- Produce their own weights… 2 (1%)
- Document how weights were determined… 1 (0.6%)

Note: Just for dictionary-based CATA research, but I suspect a broader search of management content analysis research would yield similar numbers.
Current state of the literature

- Uniform term weighting
  - All words count equally
  - …but should they?

- Why?
  - Institutionalized
  - Easy/convenient
  - How to weight?
  - Theory should drive methods
So how do we weight? (Manual)

- Can end up a lot like a survey

- Semantic differential: **How socially oriented is the author of this text?**
  Prosocial |____|____|____|____|____|____|____| Antisocial

- Likert scale: **The author of this text is socially oriented.**
  Strongly agree | Agree | Don’t Know | Disagree | Strongly Disagree
So how do we weight?
(Dictionary-based CATA: Individual words)

- Unclear… so let’s look at options

- Term Frequency-Inverse Document Frequency (TF-IDF)
  - Commonly used in Information Retrieval (e.g., Google search)
  - Words discriminate best when they:
    - Are used frequently in some texts (term frequency)
    - Are not used in all texts (inverse document frequency)

- The challenge:
  - Penalizes common-but-relevant words (‘optimistic’ vs ‘panglossian’)
  - Isn’t concerned with *polarity* (‘like’ vs ‘love’)

-
So how do we weight? (Dictionary-based CATA: Individual words)

Uses a Bayesian algorithm to assign each word a value from 0-100

Kovács et al (2013) – AllOurIdeas.org
So how do we weight?
(Dictionary-based CATA: Individual words)

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Score</th>
<th>Keyword</th>
<th>Score</th>
<th>Keyword</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Authentic</td>
<td>95</td>
<td>Truthful</td>
<td>68</td>
<td>Usual</td>
<td>53</td>
</tr>
<tr>
<td>Genuine</td>
<td>92</td>
<td>Unmistakable</td>
<td>68</td>
<td>Decent</td>
<td>51</td>
</tr>
<tr>
<td>Real</td>
<td>88</td>
<td>Artisan</td>
<td>67</td>
<td>Unusual</td>
<td>51</td>
</tr>
<tr>
<td>Skilled</td>
<td>83</td>
<td>Unpretentious</td>
<td>67</td>
<td>Caring</td>
<td>49</td>
</tr>
<tr>
<td>Faithful</td>
<td>81</td>
<td>Heartful</td>
<td>66</td>
<td>Caring</td>
<td>49</td>
</tr>
<tr>
<td>Legitimate</td>
<td>81</td>
<td>Delicious</td>
<td>65</td>
<td>Replicating</td>
<td>46</td>
</tr>
<tr>
<td>Original</td>
<td>80</td>
<td>Virtuous</td>
<td>64</td>
<td>Offbeat</td>
<td>43</td>
</tr>
<tr>
<td>Traditional</td>
<td>79</td>
<td>Normal</td>
<td>63</td>
<td>Atypical</td>
<td>41</td>
</tr>
<tr>
<td>Pure</td>
<td>78</td>
<td>Creative</td>
<td>62</td>
<td>Unassuming</td>
<td>37</td>
</tr>
<tr>
<td>Historical</td>
<td>77</td>
<td>Interesting</td>
<td>62</td>
<td>Invented</td>
<td>36</td>
</tr>
<tr>
<td>Sincere</td>
<td>77</td>
<td>Orthodox</td>
<td>62</td>
<td>New</td>
<td>36</td>
</tr>
<tr>
<td>Master chef</td>
<td>75</td>
<td>Artful</td>
<td>60</td>
<td>Unconventional</td>
<td>36</td>
</tr>
<tr>
<td>Craftsmanship</td>
<td>74</td>
<td>Special</td>
<td>60</td>
<td>Unconventional</td>
<td>36</td>
</tr>
<tr>
<td>Honest</td>
<td>74</td>
<td>Righteous</td>
<td>58</td>
<td>Peculiar</td>
<td>35</td>
</tr>
<tr>
<td>Integrity</td>
<td>74</td>
<td>Authoritative</td>
<td>56</td>
<td>Outlandish</td>
<td>32</td>
</tr>
<tr>
<td>Quintessential</td>
<td>74</td>
<td>Typical</td>
<td>56</td>
<td>Assumed</td>
<td>30</td>
</tr>
<tr>
<td>Expert</td>
<td>73</td>
<td>Awesome</td>
<td>55</td>
<td>Quirky</td>
<td>29</td>
</tr>
<tr>
<td>Iconic</td>
<td>73</td>
<td>Moral</td>
<td>55</td>
<td>Extroverted</td>
<td>28</td>
</tr>
<tr>
<td>Inspiring</td>
<td>73</td>
<td>Eccentric</td>
<td>54</td>
<td>Modern</td>
<td>27</td>
</tr>
<tr>
<td>Unique</td>
<td>72</td>
<td>Ethical</td>
<td>54</td>
<td>Unorthodox</td>
<td>27</td>
</tr>
<tr>
<td>Wholesome</td>
<td>72</td>
<td>Fresh</td>
<td>53</td>
<td>Pretentious</td>
<td>19</td>
</tr>
<tr>
<td>Professional</td>
<td>70</td>
<td>Old-fashioned</td>
<td>53</td>
<td>Untraditional</td>
<td>17</td>
</tr>
<tr>
<td>Skillful</td>
<td>70</td>
<td>Artificial</td>
<td>14</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Keywords added by participants.
So how do we weight?
(Dictionary-based CATA: Entire dictionaries)

Uses Machine Learning to assign weights to Dictionary results

<table>
<thead>
<tr>
<th>Linguistic feature</th>
<th>Source</th>
<th>Description</th>
<th>Relation with extraversion</th>
<th>Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique</td>
<td>LIWC</td>
<td>Measure of repetition of words in a given text.</td>
<td>–</td>
<td>.6457</td>
</tr>
<tr>
<td>MEANP</td>
<td>MRC</td>
<td>Paivio meaningfulness, defined as the mean value of written associations people list with a word in 30 seconds. (Paivio, 1968)</td>
<td>+</td>
<td>.3553</td>
</tr>
<tr>
<td>We</td>
<td>LIWC</td>
<td>The relative number of times the first-person plural is used, e.g., “we,” “us,” “our” (11 words).</td>
<td>+</td>
<td>.2845</td>
</tr>
<tr>
<td>T-L-FREQ</td>
<td>MRC</td>
<td>Measure of how frequently words are used in the English language. (Thomdike and Lorge, 1944)</td>
<td>–</td>
<td>.2544</td>
</tr>
<tr>
<td>Number</td>
<td>LIWC</td>
<td>The relative frequency of numbers in the text, e.g., “one,” “thirty,” “million” (29 words).</td>
<td>–</td>
<td>.2468</td>
</tr>
<tr>
<td>Motion</td>
<td>LIWC</td>
<td>The relative frequency of words related to motion in the text, e.g., “walk,” “move,” “go” (73 words).</td>
<td>+</td>
<td>.2464</td>
</tr>
<tr>
<td>Insight</td>
<td>LIWC</td>
<td>The relative frequency of words related to insight, e.g., “think,” “know,” “consider” (116 words).</td>
<td>–</td>
<td>.2355</td>
</tr>
<tr>
<td>Up</td>
<td>LIWC</td>
<td>The relative frequency of words like “up,” “above,” “over” (12 words).</td>
<td>–</td>
<td>.2296</td>
</tr>
<tr>
<td>NLET</td>
<td>MRC</td>
<td>Average number of letters in a word.</td>
<td>–</td>
<td>.2282</td>
</tr>
<tr>
<td>WPS</td>
<td>LIWC</td>
<td>Average number of words per sentence.</td>
<td>+</td>
<td>.2219</td>
</tr>
</tbody>
</table>
Project In-Progress
(Alphabetical)

Jason Kiley
Tim Michaelis
Clay Posey
Project In-Progress

- Comparison of term weighting approaches

- Existing approaches
  - Unweighted
  - TF-IDF
  - AllOurIdeas.org

- New approaches
  - Machine Learning
  - Item Response Theory
Machine Learning
Item Response Theory: Discrimination

Probability of using a word (e.g., “love”)

Your “true” score (e.g., positive affect)
Item Response Theory: Difficulty

Probability of using a word (e.g., “love”)

Your “true” score (e.g., positive affect)
Questions?