Literary Analysis Tool: Text Analytics for Creative Writers

Austin Grimsman

Douglas M. Kline
klined@uncw.edu
Information Systems

Ron Vetter
vetterr@uncw.edu
Computer Science

Curry Guinn
guinnc@uncw.edu
Computer Science

University of North Carolina Wilmington
Wilmington, NC 28403, USA

Abstract

Creative writers struggle with obtaining reliable and consistent readers for their draft works. Human reviewers are notoriously inconsistent across different reviewers, and a single reviewer's feedback can vary significantly over time. Additionally, there are logistical issues with human feedback. We apply text analytics techniques to literary works with the goal of aiding writers in revisions. For a set of text, the Literary Analysis Tool (LAT) provides objective statistics, windowed statistics over the length of the text, and mood analysis. The LAT provides quick feedback, at any time, in an absolutely objective manner. We present the feedback of a small set of creative writers. The results indicate that text analytics has a place in the creative writing process.

Keywords: Creative writing, literary analysis, text analytics.

1. INTRODUCTION

Creative writers typically follow this general cyclical process: draft, get review, revise, get review, revise… Reviews are necessary to make sure that what the author intends to convey is what they actually conveyed.

Historically, reviews are performed by a human, preferably not the author, who lacks the distance necessary to be objective. However, human reviewers have disadvantages. They are subject to personal biases & moods resulting in inconsistent and conflicting feedback. Furthermore, performing a good review is time consuming, mentally demanding, and requires patience, enthusiasm, and experience. As a result, good reviewers are hard to find, and it may take a long time to get a review.

We present the Literary Analysis Tool (LAT) which attempts to give useful feedback and addresses some of the disadvantages of human reviewers. The LAT uses standard text analytics techniques applied to the unique domain of creative writing. The LAT has the advantages of being on-demand

and perfectly objective, but cannot realistically compare to the experience and nuanced analysis of a human reviewer. Furthermore, a computational analysis might ease the burden on human reviewers by analyzing tedious items such as punctuation, length of sentences, and indirect language.

Text Analytics

The LAT uses traditional statistical based text analytics described by (Salton, Automatic Text Processing, 1989) and (Salton & McGill, Introduction to modern information retrieval, 1983). This type of analysis is essentially based on word frequencies, with the supposition that higher frequencies indicate importance.

This type of analytics can be improved by counting words with a known purpose or meaning. For example, we might count the occurrences of words in this set {death, dead, dying, mortality}. A high count might indicate a focus on death and general dark tone to a passage.

The Linguistic Inquiry and Word Count (LIWC) application takes this approach (Pennebaker, Booth, & Francis, 2007). The LWIC application has a large number of word lists, categorized by linguistic purpose (pronouns, verbs, prepositions etc.) as well as concepts such as "Family", "Anger", "Time", "Sexuality", "Death", etc.

LIWC also counts:

- First person singular and plural pronouns
- Third person singular and plural pronouns
- Parts of speech: articles, nouns, verbs, prepositions, conjunctions, negations, etc.
- Past, present, and future tenses.

Less astonishing analytics can also be performed that are very helpful for a creative writer. Metrics such as word counts and spelling check are now standard in word processors. Additional metrics that are helpful to a creative writer include: length of paragraph, frequency of words, counts of punctuation usage, and words repeated within a window, e.g., the word "really" was used twice within 5 words.

In this exploratory study, the LAT does not implement more sophisticated methods such as Natural Language Processing or deep learning models. If the easily-implementable statistical methods prove useful to writers, more sophisticated measures can be pursued.

2. PROBLEM STATEMENT

Writing is a nebulous creative process. Good finished products can be accomplished individually, but external feedback is acknowledged as beneficial. A writer's perception of their work is influenced by their journey in writing. External review is the only real way for a writer to know what they have written, rather than what they think they wrote.

Writers have traditionally relied on human reviewers. However, humans bring their own personal biases and moods to the reading. Different reviewers can provide widely variable and often opposite feedback. Relying on a single reviewer might result in positive / negative, so multiple reviews are advisable.

Unfortunately, reviewing is difficult and time consuming. It takes substantial mental effort for a focused reading with the added cognitive load of assessing on many dimensions what is being read. Good reviewers should be experienced, honest, and willing to spend the time to perform the review.

For the writer, obtaining external reviews is an external dependency over which they have little control. Overly harsh/soft feedback from an overly depressed/happy reviewer can significantly change the direction of revisions. A long delay in a review can also impact the writing process.

In short, humans are not ideal reviewers. Attributes of a truly ideal reviewer include:

- Instant feedback, on demand
- Consistency – same feedback for same work
- Feedback on technicalities: punctuation, word counts, etc.
- Feedback on style: passive voice, flowery language
- Feedback on mood
- Feedback over the entire passage, as well as the throughout the reading

The LAT system is an attempt to address these issues using straightforward text analysis techniques.

3. LAT SYSTEM

The goal of the LAT system is to assess the usefulness of standard text analytics for creative writers. Considering this goal, the choice was made to forego application integration with word processors. Application integration (with, for example MS Word) would have been dependent on particular word processors and versions, and

would have required deployment across users machines.

HTML, CSS, and Javascript proved to be a good prototyping platform, and the entire system was deployed as a single page web site. Deployment was a matter of sending a URL to a user. Appendix A gives a screenshot of the LAT system. Users paste a passage on the left-hand textbox, then choose an analysis button. Results are shown in the right-hand textbox. Graphical representations are shown below the buttons and can be seen by scrolling down to them.

To explore the utility of text analytics for creative writers, the LAT implemented very standard, easily implemented metrics:

- LIWC Analysis
- Language Tone Analysis
- Paragraph check
- Word Frequency
- Punctuation
- Indirect Language
- Word Proximity
- Moving Window Language Tone

Table 1: LIWC parts-of-speech analysis for Poe's "The Cask of Amontilado"

Word Count: 2348
Function Words: 1332
Total pronouns: 359
Personal pronouns: 261
1st pers singular pronouns: 158
1st pers plural pronouns: 24
2nd pers pronouns: 33
3rd pers singular pronouns: 86
3rd pers plural pronouns: 13
Impersonal pronouns: 98
Articles: 240
Common verbs: 271
Auxiliary verbs: 171
Past tense: 142
Present tense: 81
Future tense: 25
Adverbs: 71
Prepositions: 317
Conjunctions: 121
Negations: 46
Quantifiers: 41
Numbers: 35
Swear words: 0

LIWC Analysis

LIWC analysis provides counts of words in many categories. The categories include parts-of-speech, as well as general categories of meaning. Parts-of-speech output is a count of the number of instances of, for example: number of words, total number of pronouns, 1st person singular pronouns, 3rd person plural pronouns, impersonal pronouns, verbs, past tense verbs, adverbs, prepositions, conjunctions, quantifiers, negations, numbers, and swear words.

Table 2: LIWC Concept Analysis for Poe's "The Cask of Amontilado"

Social processes: 235
Family: 2
Friends: 8
Humans: 8
Affective processes: 142
Positive emotion: 77
Negative emotion: 65
Anxiety: 13
Anger: 13
Sadness: 14
Cognitive processes: 265
Insight: 32
Causation: 20
Discrepancy: 17
Tentative: 21
Certainty: 21
Inhibition: 9
Inclusive: 108
Exclusive: 45
Perceptual processes: 63
Seeing: 12
Hearing: 34
Feeling: 14
Biological processes: 62
Body: 37
Health: 12
Sexual: 3
Ingestion: 12
Relativity: 302
Motion: 50
Space: 151
Time: 103
Work: 14
Achievement: 24
Leisure: 13
Home: 5
Money: 7
Religion: 2
Death: 9
Assenting: 10
Nonfluent words (er, umm): 3
Filler words: 2

Where LIWC really shines is in relating words to abstract concepts and emotions. For example LIWC provides a count of words in the passage that relate to family, such as {kin, mom, dad, husband, son, family, uncle, relative, etc.}, and variations on those. As another example, LIWC provides a count of words related to anger, such as {agitate, angry, bitter, cruel, destroy, fury,

jerk, mad, jealous, revenge, etc.} Details on how these word lists were developed can be found here (Tausczik & Pennebaker, 2010).

Language Tone Analysis

One drawback of the LIWC analysis is the amount of detail – LIWC breaks down to 64 different constructs. To address this, we provide a more aggregated, and faster executing, analysis.

Table 3: Language Tone output for Edgar Allen Poe's "The Cask of Amontilado"

Word Count: 2348
Positive Emotions: 77
Negative Emotions: 65
Total Emotional Intensity: 142
Cognitive Mechanisms: 265
Motive Concerns: 63
Perceptual/Personal Processes: 116

Paragraph, Frequency, Punctuation, and Indirect Language, Word Proximity

Tables 4 through 7 give the output for the LAT's analysis of Paragraphs, Word Frequency, Punctuation and Indirect Language, respectively.

Table 4: Paragraph Analysis

Paragraphs: 89
Words: 2348
Avg Paragraph length: 26.4

Table 5: Word Frequency

said: 24
amontillado: 16
upon: 15
ugh: 15
fortunato: 14
will: 13
us: 10
one: 8
replied: 8
let: 8
friend: 7
yes: 7
luchresi: 6
go: 6
back: 6
long: 6
catacombs: 6
bones: 6
must: 5
<shortened for space>

Depending on the audience and the writer's intention, longer or shorter paragraphs may be desirable. Feedback on word frequency may be a sign to the writer of the actual emphasis on certain topics. Table 6 shows punctuation, and it is interesting to note that Poe used 28 semicolons in his short story The Cask of Amontillado. Table 7 shows counts of indirect language. The number of adverbs is one imperfect measure. Examples of indirect language are {generally, commonly, presumably, could, might, etc.}.

Table 6: Punctuation

Periods: 177
Commas: 158
Colons: 0
Semicolons: 28
Apostrophes: 3
Quotation Marks: 166
Exclamation Marks: 49
Brackets: 0
Parentheses: 0
Braces: 0
Hyphens: 36
Ellipses: 0
Em Dashes: 29

Table 7: Indirect Language

9 indirect language phrases.
71 adverbs.

Table 8: Word Proximity, 5 word window

the: 162
he: 79
ugh: 73
i: 50
of: 22
yes: 18
ha: 13
a: 10
true: 8
in: 8
and: 8
punish: 6
you: 6
as: 6
it: 6
mason: 6
are: 4
to: 4
not: 4
tell: 2
will: 2
sign: 2
with: 2

Table 8 shows the output for a five-word window. As an example, the word "ugh" 73 times in a moving window of five words throughout the text. Changing the window to 2, shows that "ugh" appears 28 times directly next to itself, specifically, in the passage that includes 15 instances in succession.

Windowed Tone Analysis

The LAT provides a way to view the dynamics of a passage from beginning to end. This is implemented as a windowed analysis where the window is adjustable. Appendix B gives an example by displaying a count of positive and negative emotion words with a window size of 500. The x-axis represents the word location in the passage, the blue line represents frequency (count) of negative-emotion words, and the tan line represents frequency of positive-emotion words. The window is adjusted at the beginning and end, for example, the window at word 1 is 250 words, representing the first 250 words, the window at word 2 is the first 251 words, with the first full 500-word window occurring at word 250.

For those familiar with the short story The Cask of Amontillado, the negative emotions surpass the positive at the point when Fortunado endures a fit of coughing, and the narrative becomes immediately darker with words including {health, ill, kill, die, caution, buried, repose, serpent, fangs, skeletons, etc.}

Appendix C displays a similar graph, this time displaying the frequency of perceptual words exemplified by colors, smells, sounds, touch, and body parts (especially sensory organs). The frequency of perceptual words again makes sense for this short story, building up through exposition and as the characters move into the crypt, then descending through dialog and action.

4. RESULTS

To assess the utility of the LAT, we performed a small study by enlisting a convenience sample of creative writers. The enlisted writers self-assessed their experience level, and ranged from inexperienced to published experience writers. Nine candidates were invited, with a clear expectation that the assessment would require at least an hour of their time, and also require learning the relatively simple LAT software. Six ultimately completed the assessment. It took approximately six weeks for the six writers to complete the assessment.

Efforts were made to not bias the writers, but interactions were inevitable, mainly in helping use the LAT. Writers were instructed to try the LAT's various analyses on existing works that they knew well, and also their own work. Free-form experimentation with the LAT was to last at least 30 minutes, after which they could complete a short (12 question survey).

Respondents rate the various analysis features on a 5-point likert scale where 1 was "Not Useful" and 5 was "Extremely Useful". Raters varied in their overall opinions, with the average across-the-board ratings mean ranging from 2.11 for one rater to 4.11 for another. The results are shown in Table 9. Due to the limited sample size, all ratings are shown, in ascending order, along with the mean rating.

Table 9: Writer Assessments of the LAT

feature	Ratings	Mean
Language Tone	2, 3, 3, 3, 4, 5	3.33
LIWC Analysis	1, 2, 3, 4, 4, 5	3.17
Paragraph Check	2, 2, 2, 3, 3, 5	2.83
Word Frequency	2, 2, 3, 3, 5, 5	3.33
Punctuation	1, 2, 2, 3, 4, 4	2.67
Indirect Language	2, 2, 3, 3, 4, 4	3.00
Word Proximity	2, 2, 3, 4, 4, 5	3.33
Overall Usefulness	2, 2, 3, 4, 4, 5	3.33
Would you seriously consider using the Literary Analysis Tool to assist in your creative writing process?	4 – Yes 2 – No	
Would you seriously consider using the Literary Analysis Tool to assist your writing process if it were further refined in the future?	6 – Yes 0 - No	

5. CONCLUSION

The last question from the survey was encouraging, in that writers found at least some potential for a useful system. The most highly rated features were the Language Tone Analysis, the Word Frequency, and the Word Proximity report. The lower rated features included the Punctuation report and the Paragraph report.

The results must be moderated by consideration of the low sample, and the intangible, subjective nature of rating "usefulness to the writing process". However, the prototyping process has appeared to clearly work, indicating that a more polished application merits consideration.

We also received informal feedback regarding absolute versus relative metrics. For example,

knowing that the average paragraph length is 65 words, an absolute metric, is difficult to interpret and take action on. In contrast, seeing a graph of the number of positive-emotion words is more helpful, in that it compares a point in the passage with previous and following narrative. Perhaps, a metric such as paragraph length would be better presented as a relative metric, pointing out paragraphs that are unusually small or large.

In the future, a more polished system might be integrated into the users preferred word processor, with reports available through a context-sensitive menu for a particular selection of text. This would certainly avoid the cut-and-paste necessary steps necessary for the current LAT. With this improvement, writers might more fully integrate feedback solicitation into their writing process. It could be that the extra steps necessary did not allow writers to take advantage of the on-demand, immediate benefits of an automated analysis.

6. REFERENCES

Adamic, L. A. (2000). Zipf, power-laws, and pareto-a ranking tutorial. Palo Alto, CA: Xerox Palo Alto Research Center.

Ananiadou, S. (2009). Adding Value to Scholarly Communications & Repositories through Text Mining. Manchester, UK: The University of Manchester. Retrieved 1 1, 2013, from https://indico.cern.ch/event/48321/contributions/1992204/attachments/957243/1358661/OAI6_SA.pdf

Bartell, B. T., Cottrell, G. W., & Belew, R. K. (1995). Representing documents using an explicit model of their similarities. Journal of the American Society for Information Science, 46(4), 254-271.

Charniak, E. (1996). Statistial Language Learning. MIT Press.

Clifton, C., Cooley, R., & Rennie, J. (2004). TopCat: data mining for topic identification in a text corpus. IEEE Transactions on Knowledge and Data Engineering, 16(8), 949-964.

Hespos, S. J., & Spelke, E. S. (2004). Conceptual precursors to language. nature: international journal of science, 430(6998), 453-456. doi:10.1038/nature02634

Mischne, G. (2007). Applied text analytics for blogs. Universiteit van Amsterdam. Retrieved from https://pure.uva.nl/ws/files/4378014/47196_mishne_thesis.pdf

Pang, B., & Lee, L. (2008). Opinon Mining and sentiment analysis. Foundations and Trends in Information Retrieval 2, 1-135.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Operator's Manual Linguistic Inquiry and Word Count: LIWC2007. Austin, TX: The University of Texas at Austin and The University of Auckland, New Zealand.

Ramsay, S. (2003). Reconceiving Text Analysis: Toward an Algorithmic Criticism. Literary and Linguistic Computing, 18(2), 167-174. Retrieved from https://doi.org/10.1093/llc/18.2.167

Salton, G. (1989). Automatic Text Processing. Reading: Addison-Wesley Publishing Company.

Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval. McGraw-Hill.

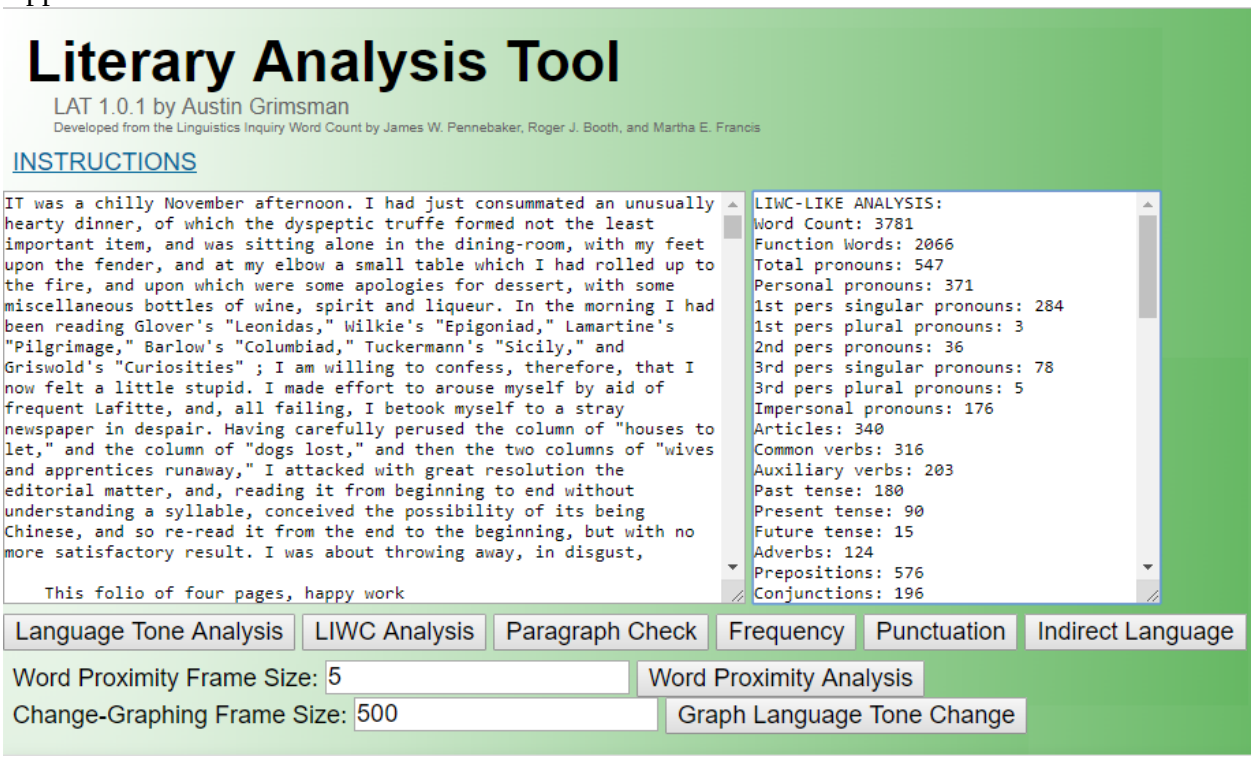
Shutova, E. (2010). Models of metaphor in NLP. ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, (pp. 688-697). Uppsala, Sweden. Retrieved from https://dl.acm.org/citation.cfm?id=1858752

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology, 29(1), 24-54.

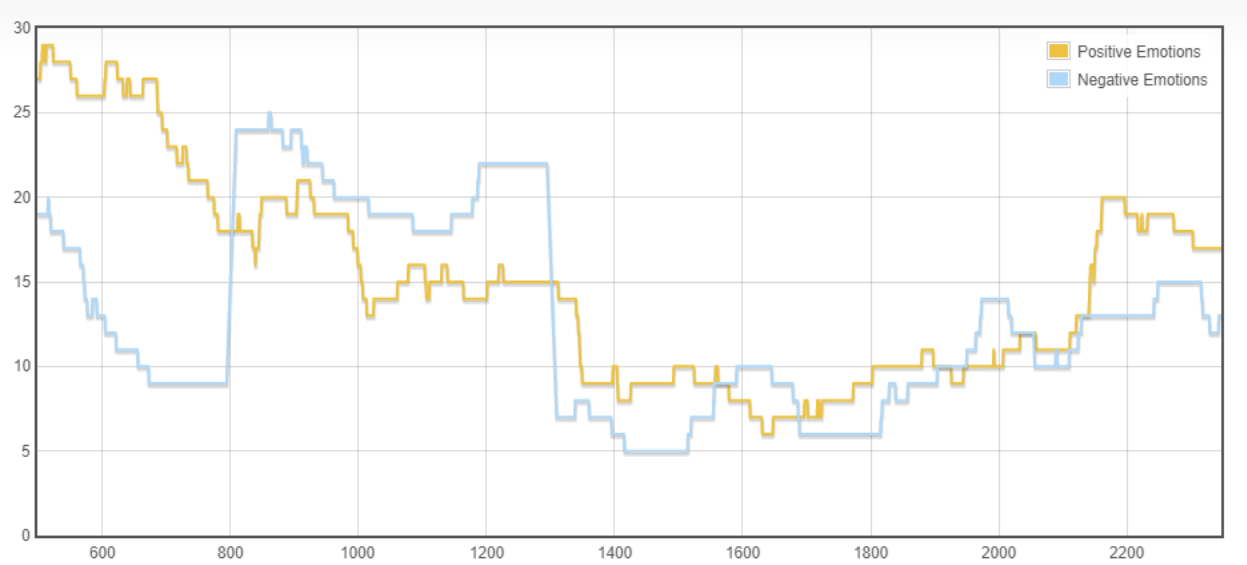
Williams-Whitney, D., Mio, J. S., & Whitney, P. (1992). Metaphor production in creative writing. Journal of Psycholinguistic Research, 21(6), 497-509.

Appendices and Annexures

Appendix A: LAT user interface



Appendix B: Windowed Analysis of Positive and Negative Emotions



Appendix C: Windowed Analysis of Perceptual Words