CS 424 - RESEARCH ARTICLE SUMMARY

For this assignment, I chose the paper titled “From Once Upon a Time to Happily Ever After: Tracking Emotions in Novels and Fairy Tales” by Saif Mohammad, National Research Council, Canada.

After reading the paper I present my critique of the paper below, based on the questions given in the assignment -

CONTEXT

Reading the title we can conclude that this paper analyzes texts from literary books, specifically novels and fairy tales, which are the primary focus. As the author notes, due to widespread digitization of text, we have unprecedented amount of literary text available to us that can be used for various analytical tasks. This paper shows how we can use sentiment analysis of text together with visualization to quantify and track emotions in individual books and large collections.
The author explores the problem that though we have huge amount of data, access and analysis techniques rely majorly on keyword searches alone. This paper shows that quantifying and tracking emotions can serve many purposes such as allowing searches based on emotions, e.g. finding snippets from the Sherlock Holmes series that build the highest sense of anticipation and suspense; identifying how books have portrayed different people and entities over time; performing comparative analysis of literary works, genres and writing styles; automatically generate summaries that captures different emotional states of the characters in a novel and analyzing emotion words and their role in persuasion.
Exploring this problem is important because literary texts have long been channels to convey emotions. A simple example is that if readers are searching for a suspense story, instead of doing a keyword search and going through a list of random results, they can be shown list of stories with guaranteed suspense with high emotion density - a concept introduced in this paper. This not only affects literary researchers but also general public. Literary researchers as well as casual readers may be interested in noting how the use of emotion word has varied through the course of a novel. Everyone interested in reading or analyzing literary work will benefit from reading this paper as this paper introduces a concept that will give them deeper understanding of the text they read.

CONTRIBUTIONS

This paper introduces the concept of emotion word density to show how collections of text can be organized for better search. Also it shows that how sentiment analysis can be used in tandem with effective visualizations to quantify and track emotions in individual books or large collections.It also presents a comparison of emotion words in novels and fairy tales. It's an interesting discovery that fairy tales have a much wider range of emotion word density than novels.

Visualizations of emotion words in Shakespeare's work i.e. Hamlet and As You Like It, has been done to show the difference between two novels by showing the difference between percentage scores of each of the emotions. It is concluded that Hamlet has more fear, disgust, sadness and anger and less joy, trust and anticipation. This type of explicit analysis is very intriguing. This paper also shows that by analyzing various segments of a novel, we can trace the flow of emotions in that novel. For example, it is found that the novel Frankenstein is much darker in the final chapters.

STRUCTURE

The paper is very well structured. It has been divided into six sections. The first section i.e. Introduction explains the purpose of the study and its direct and indirect implementations. Also in that section it has been explained how the rest of the paper is structured. The author's work is part of a broader project to provide an affect based interface to Project Gutenberg. The author explains his data in Section three and how it was collected and the steps that were taken. Section four use visualizations to analyze the texts of famous literary work of Shakespeare, Hamlet and As You like it. Various conclusions are drawn after performing emotion word analysis on the texts of both novels. Bar charts, word clouds and line graphs are used to quantify, emphasize and analyze the emotion words in these novels. The steps taken are introducing the task, explaining the data, visualizing the data and analyzing the data.

RESULTS

The paper presented an emotion analyzer which was based on the word-emotion association lexicon. The use of emotion words in individual texts and large collections was tracked and analyzed using a number of visualizations. The paper introduced the concept of emotion word density and using Brothers Grimm fairy tales as example, to show how large collections of text can be organized for better search. Another result was to show how to determine emotion association portrayed in books towards different entities which is interesting because it also gives the reason behind the emotion by looking at the world history. For e.g. The mention of fear words in close proximity of Germany spiked during the world war 2 era in texts.
Specific conclusions are made about novels and fairy tales that fairy tales have a much wider distribution of emotion word density than novels. This helps the target audience better understand the emotional content and flow of emotions in the texts.

METHODS

All the work in this paper, all the conclusions are based on the data derived by mining the content of books published over significant period of time. Data was used from various sources such as NRC Emotion Lexicon, Google N-gram corpus, Corpus of English Novels (CEN), Fairy Tale Corpus (FTC), Brothers Grimm fairy tales and Shakespeare's novels. There was no user study or survey or interview involved. The visualizations does present a statistical analysis of data which was precalculated. The methods used are very thorough and descriptive. However,

the author has assumed the accuracy of the google corpus data and that of the NRC lexicon to achieve his results. The NRC lexicon is prone to human error and similarly google's data is prone to computer generated errors. The addition of more data would definitely help achieve more finer results and possibly a saturation point where addition of more data does not affect the result.

TECHNOLOGIES

The author does not explore any specific device or technology rather uses digitized version of texts along with text mining and statistical analysis algorithm to compute the results. It is hard to comment at this point that there could be a device that could make use of this paper but if I had to imagine I would say a future device that analyzes people's conversations by observing and collecting and emotionally analyzing it. But as it has been noted in this paper the technology is still developing and can be unpredictable for shorter sentences.

FIGURES

The visualizations used in this paper are bar charts, word clouds and scatter plots. There are 20 figures in this paper which depict the result of the experiments discussed in this paper. Figure 1 and 2 show the percentages of emotion words in Shakespeare's famous tragedy, Hamlet and his comedy, As you like it, respectively. Figure 3 shows the difference in the percentages which helps explain the difference in the two novels in terms of emotion words and captures the main idea behind the paper. Word clouds capture relative salience of trust words such as brother, marry etc and sadness words such as death, late etc between the two novels. Another idea was to show emotion distribution of snippets of texts to enable search by emotion and Figure 6-8 captures exactly how that can be achieved. In those figures timeline of emotions is shown evolving over time in twenty segments.
The concept of emotion word density is conveyed in figure 9 -11. Figure 9, a line graph, is interactive and a user can select two stories to perform more detailed analysis between the selections. Figure 12-14 capture emotions associated with targets.Figure 15 -20 are histograms comparing novels and fairy tales for positive and negative emotions.
The figures are simple yet they convey the desired meaning clearly. In terms of user experience, the figures are easy to understand.

CONFUSING WORDS OR IDEAS

There wasn't anything specifically confusing for me in this paper because I do have some NLP (Natural Language Processing) background. However, for some people the words like lexicon, N-grams can be hard to understand. Lexicon is the vocabulary of a person or language, or a branch of knowledge and an **n**-**gram** is a contiguous sequence of **n** items from a given sequence of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. The **n**-**grams** typically are collected from a text or speech corpus. The idea behind creating the NRC lexicon is a little confusing to me. How does the author justify

the use of eleven questions for determining word-emotion association. The use of only five people to annotate text also feels less. Using more people could help in creation of a more accurate word-emotion lexicon. Around ten thousand entries of word-sense pair are discarded which could be utilized by further analysis.

EVALUATION

The paper effectively compares a collection of novels and a collection of fairy tales and draws important conclusion in terms of emotion word density - a concept introduced in this paper. This work, as he author notes, is reliable for analyzing large volumes of text and thus can be applied to any literary work. Further improvements can be made by improving the lexicon that captures the word-emotion association more accurately. The use of 5-grams in this paper can be extended to 6-grams or higher to see if the results differ and if they do then by what margin. The future work as noted by the author himself is to provide an affect-based interface to Project Gutenberg. Further, the users will be able to search for snippets from multiple texts that have strong emotion word densities. This paper has been cited 64 times (information provided by google scholar). Some examples are - Sentiment Analysis and Opinion Mining: Synthesis Lectures on Human Language Technologies by Bing Liu, UIC; EmpaTweet: Annotating and Detecting Emotions on Twitter, K Roberts et. al.

RELEVANCE TO OUR PROJECT

We are creating a visualization in the same domain as this paper. Earlier, we has also decided to do some sort of sentiment analysis on books data and after reading this paper our goals have become more clear. This is definitely a technique that we can incorporate in our project by performing the same sort of emotion word analysis of the books we intend to use in our project. This paper was written five years ago and the latest versions of the data banks could be more reliable now.

LINKS

Citations List Link:
https://scholar.google.com/scholar?cites=12220666054977470082&as_sdt=400005&sciodt=0,14&hl=en

Published in:
Proceeding
LaTeCH '11 Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities

Link to the paper:
http://dl.acm.org/citation.cfm?id=2107650