Decision Support Systems 53 (2012) 730–741

Contents lists available at SciVerse ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

From once upon a time to happily ever after: Tracking emotions in mail and books

Saif M. Mohammad

Emerging Technologies, National Research Council Canada , Ottawa, Ontario, Canada, K1A 0R6

ARTICLE INFO

Available online 23 May 2012

Keywords:
Emotion analysis
Sentiment analysis
Lexicon
Email
Fairy tales
Enron Corpus
Google Books Corpus

ABSTRACT

In this paper, we show how sentiment analysis can be used in tandem with effective visualizations to quantify and track emotions in mail and books. We study a number of specific datasets and show, among other things, how collections of texts can be organized for affect-based search and how books portray different entities through co-occurring emotion words. Analysis of the Enron Email Corpus reveals that there are marked differences across genders in how they use emotion words in work-place email. Finally, we show that fairy tales have more extreme emotion densities than novels.

1. Introduction

Emotions are an integral part of how humans perceive and communicate with the outside world. We convey emotion through our facial expressions, our speech, and through our writing. A given sentence may be pertinent to many different entities and determining the emotions evoked by one entity in another is fairly challenging—often requiring information not present in the sentence itself. For example, consider this headline in a newspaper from 2009:

When your cartoon can get you killed.

The article is about the controversy surrounding a particular episode of the television series *South Park*. The entities involved here are the creators of the show and the extremists issuing them death threats. The sentence has a writer and a large number of readers. All of these people may be expressing or feeling certain emotions. However, identifying the emotions associated with different entities requires not just the analysis of the target sentence, but often also of the context, entity behavior, and world knowledge. Thus, it is not surprising that current methods that attempt this task have relatively low accuracies [48].

However, for the first time in our history, we now have access to hundreds of thousands of digitized mail, books, and social-media communications. Even though making accurate predictions of individual instances may be error prone, simple methods can be used to draw reliable conclusions from many occurrences of a target entity. In this paper, we show how we created a large word–emotion association lexicon by crowdsourcing (Section 3), and use it to analyze the use of emotion words in large collections of text. Specifically, we show how sentiment analysis can be used in tandem with effective visualizations to quantify and track emotions in mail (Sections 4–6) and in books (Sections 7–11). Many of these techniques can also apply to data from other forms of communication, such as Twitter feeds.

The lexicon we created has manual annotations of a word's associations with positive polarity negative polarity, and eight emotions—joy, sadness, anger, fear, trust, disgust, surprise, anticipation. These emotions have been argued to be the eight basic and prototypical emotions [43].

1.1. Emotion analysis of mail

Letters have long been a channel to convey emotions, explicitly and implicitly, and now with the widespread usage of email, we have access to unprecedented amounts of text that we ourselves have written. Automatic analysis and tracking of emotions in mail has a number of benefits including:

1. *Decision Support Tool*: Helping physicians identify patients who have a higher likelihood of attempting suicide [29,39] The 2011 Informatics for Integrating Biology and the Bedside (i2b2) challenge by the National Center for Biomedical Computing is on detecting emotions in suicide notes.
2. *Social Analysis*: Understanding how genders communicate through work-place and personal email [7].
3. *Productivity and Self-Assessment Tool*: Tracking emotions towards people and entities, over time. For example, did a certain managerial course bring about a measurable change in one's interpersonal communication?

E-mail address: saif.mohammad@nrc-cnrc.gc.ca.

0167-9236/$ – see front matter.
doi:10.1016/j.dss.2012.05.030

4. *Health Applications*: Determining if there is a correlation between the emotional content of letters and changes in a person's social, economic, or physiological state. Sudden and persistent changes in the amount of emotion words in mail may be a sign of psychological disorder.
5. *Search*: Enabling affect-based search. For example, efforts to improve customer satisfaction can benefit by searching the received mail for snippets expressing anger [13,15].
6. *Writing Aids*: Assisting in writing emails that convey only the desired emotion, and avoiding misinterpretation [26].

In Section 4, we show comparative analyses of emotion words in love letters, hate mail, and suicide notes. This is done: (a) To determine the distribution of emotion words in these types of mail, as a first step towards more sophisticated emotion analysis (for example, in developing a depression–happiness scale for Application 1), and (b) To use these corpora as a testbed to establish that the emotion lexicon and the visualizations we propose help interpret the emotions in text. In Section 5, we analyze how men and women differ in the kinds of emotion words they use in work-place email (Application 2). Finally, in Section 6, we show how emotion analysis can be integrated with email services such as Gmail to help people track emotions in the emails they send and receive (Application 3).

1.2. Emotion analysis of books

Literary texts, such as novels, fairy tales, fables, romances, and epics tend to be rich in emotions. With widespread digitization of text, we now have easy access to large amounts of such literary texts. Project Gutenberg provides access to 34,000 books [24].[1] Google is providing n-gram sequences, and their frequencies, from more than 5.2 million digitized books, as part of the *Google Books Corpus (GBC)* [30].[2] Emotion analysis of books has many applications, including:

1. *Search*: Allowing search based on emotions. For example, retrieving the darkest of the Brothers Grimm fairy tales, or finding snippets from the Sherlock Holmes series that build the highest sense of anticipation and suspense.
2. *Social Analysis*: Identifying how books have portrayed different people and entities over time. For example, the distribution of emotion words used in proximity to mentions of women, race, and homosexuals.
3. *Comparative analysis of literary works, genres, and writing styles*: For example, is the distribution of emotion words in fairy tales significantly different from that in novels? Do women authors use a different distribution of emotion words than their male counterparts? Did Hans C. Andersen use emotion words differently than Beatrix Potter?
4. *Summarization*: For example, automatically generating summaries that capture the different emotional states of the characters in a novel.
5. *Analyzing Persuasion Tactics*: Analyzing emotion words and their role in persuasion [4,28].

We present a number of visualizations that help track and analyze the use of emotion words in individual texts and across very large collections, which is especially useful in Applications 1, 2, and 3 described above (Sections 7 and 8). Using the Google Books Corpus we show how to determine emotion associations portrayed in books towards different entities (Section 9). We introduce the concept of emotion word density, and using the Brothers Grimm fairy tales as an example, we show how collections of text can be organized for better search (Section 10). Finally, for the first time, we compare a collection of novels and a collection of fairy tales using an emotion lexicon to show that fairy tales have a much wider distribution of emotion word densities than novels (Section 11).

This work is part of a broader project to provide an affect-based interface to Project Gutenberg. Given a search query, the goal is to provide users with relevant visualizations presented in this paper, and the ability to search for text snippets that have high emotion word densities.

2. Related work

Over the last decade, there has been considerable work in sentiment analysis, especially in determining whether a term has a positive or negative polarity [25,34,54]. There is also work in more sophisticated aspects of sentiment, for example, in detecting emotions such as anger, joy, sadness, fear, surprise, and disgust [2,5,33]. The technology is still developing and it can be unpredictable when dealing with short sentences, but it has been shown to be reliable when drawing conclusions from large amounts of text [14,37,41].

Automatically analyzing affect in emails has primarily been done for automatic gender identification [9,10], but it has relied mostly on surface features such as exclamations and very small emotion lexicons. The WordNet Affect Lexicon (WAL) [49] has a few hundred words annotated with associations to a number of affect categories including the six Ekman emotions (joy, sadness, anger, fear, disgust, and surprise).[3] General Inquirer (GI) [47] has 11,788 words labeled with 182 categories of word tags, including positive and negative polarity.[4] Affective Norms for English Words (ANEW) has pleasure (happy–unhappy), arousal (excited–calm), and dominance (controlled–in control) ratings for 1034 words.[5] Mohammad and Turney [33] compiled emotion annotations for about 4000 words with eight emotions (six of Ekman, trust, and anticipation).

Empirical assessment of emotions in literary texts has sometimes relied on human annotation of the texts, but this has restricted the number of texts analyzed. For example, Alm and Sproat [1] annotated twenty two Brothers Grimm fairy tales to show that fairy tales often began with a neutral sentence and ended with a happy sentence. Here we use out-of-context word–emotion associations and analyze individual texts to very large collections.

Automatic systems for analyzing emotional content of text follow many different approaches: a number of these systems look for specific emotion denoting words [18], some determine the tendency of terms to co-occur with seed words whose emotions are known [45], some use hand-coded rules [35,36], and some use machine learning and a number of emotion features, including emotion denoting words [2,3].

Much recent work focuses on six emotions studied by Ekman [17]. These emotions—joy, sadness, anger, fear, disgust, and surprise—are a subset of the eight proposed by Plutchik [43]. There is less work on complex emotions, for example, work by Pear et al. [42] which focuses on politeness, rudeness, embarrassment, formality, persuasion, deception, confidence, and disbelief. Francisco and Gervás [19] marked sentences in fairy tales with tags for pleasantness, activation, and dominance, using lexicons of words associated with the three categories.

There has also been some interesting work in visualizing emotions [21,44,50]. Mohammad [31] describes work on identifying colors associated with emotion words.

[1] Project Gutenberg: http://www.gutenberg.org.
[2] *GBC*: http://ngrams.googlelabs.com/datasets.
[3] WAL: http://wndomains.fbk.eu/wnaffect.html.
[4] GI: http://www.wjh.harvard.edu/~inquirer.
[5] ANEW: http://csea.phhp.ufl.edu/media/anewmessage.html.