ROBUSTNESS OF EMOTIONS EXTRACTION FROM 20TH CENTURY ENGLISH BOOKS

Alberto Acerbi, Vasileios Lampos, R. Alexander Bentley
Abstract

We report here trends in the usage of "mood" words, that is, words carrying emotional content, in 20th century English language books, using the data set provided by Google that includes word frequencies in roughly 4% of all books published up to the year 2008. We find evidence for distinct historical periods of positive and negative moods, underlain by a general decrease in the use of emotion-related words through time. Finally, we show that, in books, American English has become decidedly more "emotional" than British English in the last half-century, as a part of a more general increase of the stylistic divergence between the two variants of English language.
(I) 20TH CENTURY “HAPPINESS”
(2) MOOD-WORDS USE DECLINE
(3) AMERICANS MORE “EMOTIONAL”
ROBUSTNESS CHECK

- new Google Corpus (from 5M to 8M books)
- Different emotions scores (original=WNA):
  - Linguistic Inquiry and Word Count (LIWC)
  - Dodd’s et al. “Hedonometer” (HED)
- Confidence Interval Estimation
- Part-Of-Speech Analysis
20th Century “Happiness”

![Graph showing happiness trends from 1900 to 2000.](image)

- **WNA** and **LIWC** happiness scores from 1900 to 2000.
- **HED** happiness scores from 1900 to 2000.
MOOD-WORDS USE DECLINE

Year

Emotion score

1900 1920 1940 1960 1980 2000

1 2 3 4

geek*
soulmate*
ROFL
laidback
sucky
crappy
nerd*
LMAO
sweetie*
scary
AMERICANS MORE “EMOTIONAL”

American–British (Z-scores)

Year

1900 1920 1940 1960 1980 2000

WNA
LIWC
Robustness of emotion extraction from 20th century English books

Alberto Acerbi
Department of Archaeology and Anthropology
University of Bristol
Bristol, United Kingdom
Email: alberto.acerbi@gmail.com

Vasileios Lampos
Department of Computer Science
University of Sheffield
Sheffield, United Kingdom
Email: bill@lampos.net

R. Alexander Bentley
Department of Archaeology and Anthropology
University of Bristol
Bristol, United Kingdom
Email: r.a.bentley@bristol.ac.uk

Abstract—In this paper, we test the robustness of emotion extraction from English language books published in the 20th century. Our analysis is performed on a sample of the 8 million digitized books available in the Google Books Ngram corpus by applying three independent emotion detection tools: WordNet Affect, Linguistic Inquiry and Word Count, and a recently proposed ‘hedonometer’ method. We also assess the statistical robustness of the extracted patterns as well as their outputs on specific parts of speech. The analysis confirms three main results: the existence of recognizable periods of positive and negative ‘literary affect’ from 1900 to 2000, a general decrease in the usage of emotion-related words in printed books that lasts at least until the 1980s, and, finally, a divergence between American and British books, with the former using more emotion-related words from the 1960s.

I. INTRODUCTION

The study of cultural dynamics has recently been transformed by the availability, and by the relative ease of storage and analysis, of massive amounts of data. Novel forms of human-generated input (e.g., Facebook, Twitter, blogs) are produced daily, forming an interesting information source for studies on social and cultural behavior. At the same time, an increasing amount of ‘traditional’ data, such as books and newspaper articles, is digitized and made available for quantitative analysis.

One criticism targeting the use of ‘Big Data’ in social and human sciences is that the vast majority of works focuses on a short-time scale. However, ‘Long Data’ [1] with a temporal range of centuries, has provided a wealth of information, which can be used to study the evolution of societies and their cultures over time.

Fig. 1: ‘Happiness’ z-scores for years 1900 to 2000 (circles show actual times series, the lines show the smoothed time series) under WNA and LIWC. Values above zero indicate generally ‘happy’ periods, and values below the zero indicate generally ‘sad’ periods.