Chapter 5
Sign of the Times:
Sentiment Analysis on Historical Text and the Implications of Language Evolution

Tyler W. Soiferman
Stevens Institute of Technology, USA

Paul J. Bracewell
DOT Loves Data, New Zealand

ABSTRACT

Natural language processing is a prevalent technique for scalably processing massive collections of documents. This branch of computer science is concerned with creating abstractions of text that summarize collections of documents in the same way humans can. This form of standardization means these summaries can be used operationally in machine learning models to describe or predict behavior in real or near real time as required. However, language evolves. This chapter demonstrates how language has evolved over time by exploring historical documents from the USA. Specifically, the change in emotion associated with key words can be aligned to major events. This research highlights the need to evaluate the stability of characteristics, including features engineered based on word elements when deploying operational models. This is an important issue to ensure that machine learning models constructed to summarize documents are monitored to ensure latent bias, or misinterpretation of outputs, is minimized.

DOI: 10.4018/978-1-7998-9426-1.ch005

INTRODUCTION

Data, as a commodity, is often touted as the new oil. Highly accessible via the World Wide Web, the written word is a type of data that is especially prevalent and potent. However, people can describe similar things with different words, writing styles and documents of varying length. Summarizing this content manually is not scalable.

Technological developments provide the ability to process massive amounts of unstructured data with the intent of automatically and consistently extracting latent patterns. Text mining is the process of transforming unstructured text into a structured format consumable within machine learning frameworks. The imposition of structure upon text then enables features to be engineered which enable meaningful patterns and new insights to be identified.

More specifically, Natural Language Processing (NLP) can be used to summarize the themes quickly and efficiently within a corpus. NLP refers to the branch of computer science concerned with creating abstractions of text that summarize collections of documents in the same way humans can. This form of standardization means these summaries can be used operationally in machine learning models to describe or predict behavior in real or near real time as required.

With the ability to summarize collection of documents at scale, there are myriad applications of this technology. As Vajjala et. al. (2020) outlined, the past decade's breakthroughs in research regarding NLP stem from increased processing power, accessibility of digitized text, as well as algorithmic enhancement to have greater generalizability and interpretability. These advancements have resulted in NLP being increasingly used in a range of diverse domains such as retail, healthcare, finance, law, marketing, human resources and many more.

A common NLP technique is sentiment analysis, which is often used to draw sentiments from text such as customer reviews or social media posts. This functionality enables businesses to efficiently analyze unstructured data that pertains to their company, leading them to conclusions about, for example, their reputation, or the overall reaction to a product.

Sentiment analysis is a family of techniques that assign polarity scores to natural language. Typically, it is treated as a supervised machine learning problem. Example sentences are supplied that have been labelled as "positive" and "negative". Given sufficient training data, learning algorithms can distinguish positive from negative language. Positive language use scores above 0.0, and negative language scores below 0.0. Importantly, digital delivery of news reporting and sports commentary provides a wealth of accurately time-stamped textual data that can be easily indexed via technological means.

Bracewell et. al. (2016) outlined a method for quantifying the collective mood of New Zealanders using mainstream online news content. Mood is quantified