Real-Time Brand Sentiment Analysis on Social Media Using Deep Learning

Jonathan Agustin Fernando Calderon Juliet Lawton

University of San Diego®

1 Introduction

Businesses increasingly leverage social media platforms to track public perception of their brands. Many use Text Classification, Sentiment Analysis, and Natural Language Processing (NLP) to extract subjective information about specific brands from the vast corpus of unstructured social media text data. We propose to construct a system that includes training, evaluating, and deploying Machine Learning (ML) models. The model will be trained to pinpoint a brand within a text and categorize sentiment towards the brand as positive, negative, or neutral. The goal is to empower businesses with real-time insights into public sentiment about their brands, enabling businesses to quickly react and adapt to shifts in public sentiment. Ultimately, the proposed system aims to enhance the ability of businesses to maintain a positive brand image and gain a competitive edge.

2 Problem Statement

Consider a set $T = \{t_1, t_2, \ldots, t_n\}$, where each element t_i represents a unique social media post. Each t_i is a textual data point with no image, video, or audio information. Define $B = \{b_1, b_2, \ldots, b_m\}$ as a finite set that symbolizes distinct brands, with b_0 as a special element signifying the absence of a recognizable brand. Define $S = \{-1, 0, 1\}$ as a finite set that represents sentiment labels. The sentiment labels are defined as follows:

$$\begin{aligned} -1 &= \text{negative sentiment} \\ 0 &= \text{neutral sentiment} \\ 1 &= \text{positive sentiment} \end{aligned}$$

The first function $f_1 : T \to B \cup \{b_0\}$ is defined as:

$$f_1(t_i) = \begin{cases} b_j & \text{if } t_i \text{ mentions brand } b_j \\ b_0 & \text{otherwise} \end{cases}$$

The second function $f_2 : B \cup \{b_0\} \times T \to S \cup \{s_0\}$ is defined as:

$$f_2(b_j, t_i) = \begin{cases} s_k & \text{if } t_i \text{ expresses sentiment } s_k \text{ towards brand } b_j \\ s_0 & \text{otherwise} \end{cases}$$

The overall task is a composition of functions:

$$f = f_2 \circ (f_1 \times id_T)$$

where $\circ$ denotes function composition and $\times$ denotes the Cartesian product. This composition maps each post to a brand (or b_0) and then assigns a sentiment label to each brand-post pair. The identity function id_T maintains the original social media post in the pair, ensuring the sentiment label is assigned to the correct post-brand pair. The output of this composition is a sentiment label for each social media post.

Limitations. The model's primary limitation is its unimodal nature. It processes only text data, ignoring non-textual elements like images, videos, and audio clips. Often found in social media posts, these elements can contain significant sentiment-related information. This information can support, contradict, or add nuance to the sentiment expressed in the text. Audio clips, for example, can provide sentiment-related

cues through tone, pitch, and other features. The proposed system, however, will identify this multimodal information and discard the input completely. Future work should consider expanding the model to a multimodal framework to provide a more comprehensive sentiment analysis of social media posts.

3 Methodology

The work is divided into four stages: assessment, development, deployment, and evaluation. Each stage informs and refines previous stages, creating a feedback loop that continuously improves the overall system.

- **Assessment.** The stage involves selecting and testing appropriate models and datasets. The specific requirements of the sentiment analysis task, the characteristics of the available data, and the nature of the problem guide this selection process. The chosen models must effectively handle the unstructured and noisy data typical of social media and accurately capture and predict sentiments. The selected datasets must accurately represent the social media environment and provide a diverse view of expressed sentiments.
- **Development.** This stage preprocesses data in real-time to simulate a live data stream and training the selected models. The unstructured and noisy nature of social media data necessitates the transformation of text into a structured format suitable for modeling. This process involves handling invalid values and discarding non-textual content such as images, videos, or audio. Automation ensures efficiency in this process. The selected models are then adapted to the preprocessed dataset and trained to identify a brand or align with the sentiment expressed in the social media text.
- **Deployment.** This stage deploys the models to process social media data in near real-time. This stage requires building infrastructure to support the models and integrating the models with social media platforms such as Twitter where data will be streamed.
- **Evaluation.** This stage measures the models' ability to identify brands and predict sentiment in social media. For a time period, social media posts are collected and manually labeled for brand and sentiment. Standard classification metrics such as Accuracy, Precision, Recall, and F1 Score are used to evaluate the models. The models are then ranked by performance.

4 Expected Behaviors & Outcomes

We expect our models to demonstrate a nuanced understanding of human language, accurately identifying brands and sentiments. Despite the inherent challenges associated with sentiment analysis and brand identification in text data, we anticipate that our models will achieve an accuracy rate exceeding 50%. We foresee our models to have real-world applicability, serving as a valuable tool for businesses to accurately analyze sentiment from social media data. The proposed system is expected to provide actionable insights that enable companies to make informed decisions.

References

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. *OpenAI Blog, 1(8)*.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Hutto, C.J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014*.