Nihad Rustamzade

SCC.413 Coursework

26/4/2024

Lancaster University

Introduction

Problem Setting

Social media platforms, such as Twitter, have become key places for people to talk and share what they think and feel every day. They are like big collections of information, showing different emotional responses to what's happening in the world, talks people are having, and the latest trends that show us how society is reacting overall. However, because the way people use words on these sites can be pretty complex and full of deeper meanings, figuring out the emotions behind tweets is a tricky task. Being able to clearly identify and sort these emotions is super important for understanding the pulse of society online and using this knowledge in various tech and communication areas.

Motivation

Understanding the affective implications of social media might yield significant consequences. Businesses benefit from improved sentiment analysis skills, which lead to better customer interactions and more focused marketing campaigns. It offers academics and politicians insights into public opinion, which may have an impact on changes to policies and public health initiatives. Enhancing emotional understanding on social media platforms may also result in richer, more interesting user experiences and settings where people feel respected and understood.

Research Question

"Can the emotional content of tweets be accurately classified, and what are the linguistic characteristics of each emotion?" is the research question that forms the basis of this work. The investigation into whether machine learning models can successfully identify and classify the complicated emotional ground that social media writings provide is guided by this question.

Contributions

This work addresses the difficult problem of text categorization based on emotional content by utilizing advanced machine learning algorithms, which advances the domains of computational linguistics and social media analytics. Two important contributions are the creation of models capable of deciphering small language clues included in tweets and a thorough examination of the many linguistic expressions of emotions on Twitter. These initiatives aim to improve algorithmic comprehension of human emotions in digital interactions while also offering useful applications and scholarly insights.

Related Work

Literature Review

The field of text sentiment analysis and text emotion detection has grown dramatically, experimenting with lexicon-based systems and sophisticated machine learning models, among other approaches. Early on in the discipline, lexicon-based approaches were mostly employed by researchers such as Pang et al. (2002), who examined texts based on the presence of emotionally charged terms that were preset. Although simple, these approaches frequently suffer from contextual issues, missing subtleties like irony or sarcasm, as demonstrated by Liu's (2010) research.

The prevalence of machine learning techniques may be attributed to their versatility and resilience, which have been made possible by advances in computer capability. Pang and Lee's (2008) research showed how Support Vector Machines (SVM) can effectively categorize the emotions expressed in movie reviews. As deep learning gained traction, its potential to analyze sentiment was further expanded. Kim (2014) and Lai et al. (2015) used convolutional and recurrent neural networks to better understand the semantic correlations found in texts.

In an effort to capitalize on the advantages of both, recent trends have moved toward hybrid models that integrate lexicon-based and machine learning techniques. Such integration was suggested by Cambria et al. (2016) as a way to enhance context awareness, and it shown to be especially successful in texts with conflicting emotions. In addition, another advancement was the incorporation of pre-trained language models like as BERT, which I also employed in my project, and GPT, as investigated by Devlin et al. (2018) and Radford et al. (2019). These models have established new standards for a range of natural language processing applications, including sentiment analysis. They were trained on large corpuses and then adjusted for particular tasks.

Even with the improvements, every strategy has its drawbacks. While machine learning approaches require large amounts of labeled data and have interpretability issues, hybrid models show promise but are difficult to apply. Lexicon-based methods also lack contextual knowledge. Even if they are strong, the latest pre-trained models are resource-intensive and may perform poorly in circumstances unique to a certain domain. This survey of the literature emphasizes the wide range of techniques used in emotional content analysis and emphasizes the continuous need to create approaches that combine scalability and sophisticated language understanding.

About the Dataset

Dataset Description

The “Emotions” dataset, that has been sourced from Kaggle, consists of many items, each of which is a text excerpt from a Twitter tweet accompanied by a label designating the main feeling conveyed. Due to the systematic classification of the emotions into six categories, this dataset is a perfect tool for a variety of computational linguistics studies. These categories provide a structured approach to analyzing emotional reactions in various social media exchanges. Numerous fields of study, such as textual analysis, sentiment analysis, and emotion categorization, will find great use in this dataset.

The other reasons for this dataset being a good fit to answer my research question includes its diverse emotional labels, tweets being sourced from real-world data, its randomness, volume and variability.

Methodology

My research uses a multi-step technique wherein the clean and standardized text is first subjected to rigorous data pretreatment to make sure it is ready for analysis. This is crucial because machine learning model performance may be greatly impacted by the caliber of data preparation.

Data Pre-processing:

Text normalization procedures are used in the first step and are essential for lowering data noise. This involves utilizing regular expressions to remove URLs, Twitter handles, hashtags, and special characters in order to reduce the dataset's sparsity and concentrate on useful textual content. In order to preserve consistency and avoid treating the same words as distinct tokens in various contexts, I also transform all text to lowercase.

Contraction Expansion:

When dealing with contractions, I extend them to their whole forms (e.g., changing "can't" to "cannot"). The capacity of the model to develop accurate associations is improved by treating all word forms consistently, which makes standardization essential for effective sentiment analysis.

Custom Handling of Non-Standard Contractions:

Conventional contraction libraries are frequently inadequate in light of the variety of text seen on social media. In order to solve this, I developed a unique method for managing contractions that entails a thorough mapping of non-standard and missing apostrophe contractions to guarantee that all potential variants are consistently extended.

Tokenization and Negation Handling:

Tokenization is the process of breaking up sentences into individual words or tokens once the text has been cleaned. Negation handling receives particular attention, as negation words and their immediate successors are handled as individual tokens. This maintains the sentiment context that negations have altered, which is essential for the sentiment analysis that follows.

Lemmatization and POS Tagging:

Part-of-speech (POS) tagging is used to classify each word into its appropriate part of speech depending on the context, prior to lemmatization. This facilitates the use of more precise lemmatization, which reduces words to their dictionary form and makes analysis easier and more efficient.

Feature Extraction:

I convert texts into a matrix of token counts using the CountVectorizer, concentrating on unigrams and bigrams to record the frequency of individual words as well as their local contexts. This approach improves the machine learning model's feature set while also helping to comprehend the dynamics of word frequency.

Sentiment Analysis:

I use two main tools for sentiment analysis: VADER, which is quick and effective in processing huge datasets, for preliminary sentiment evaluations, and BERT, which is more

sophisticated and useful in situations when sentiment nuances matter. Based on the initial sentiment scores generated by VADER, a dynamic decision is made on which of these technologies to use.

Modeling:

A logistic regression model, selected for its efficacy and efficiency in addressing binary and multi-class classification issues, forms the foundation of our investigation. The model's performance is assessed across a range of emotion categories through training and evaluation using common measures including accuracy, precision, recall, and F1-score.

Evaluation of Model Performance and Research Question Analysis

With an overall accuracy of 84.2%, our logistic regression model performed admirably, demonstrating the ability to effectively classify emotional content in tweets. Considering the difficulty of the natural language processing tasks involved, this performance is quite noteworthy. The model showed excellent recall in classes like happiness and good accuracy in classes linked with obvious emotional expressions like joy, indicating that it seldom ever misses these emotions in the dataset.

But the model struggled with more complex feelings, like dread, and its F1-scores dropped. This gap, which is probably caused by subtler language signals and fewer training instances in the dataset, emphasizes how difficult it is to strike a balance between accuracy and recall for complex emotions.

In response to the study question, the findings verify that the emotional content of tweets can be correctly categorized and that some emotions are better than others at capturing the linguistic features of each emotion. Strong performance metrics were demonstrated by emotions that were well-represented, but classes with complex emotions suggested the need for more advanced analytic methods or larger training sets.

The study's conclusions point to possible areas for development, especially in terms of strengthening the model's capacity to manage nuanced emotional displays. Subsequent research endeavours may involve the use of sophisticated machine learning models, including deep neural networks, and the enlargement of the dataset to encompass a wider range of intricate emotional expressions. This might enhance the accuracy and generalizability of the model in various emotional settings.

References:

1. Pang, B., & Lee, L. (2002). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135.
2. Liu, B. (2010). Sentiment analysis and subjectivity. Handbook of Natural Language Processing, 2, 627-666.
3. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Machine Learning, 2(1-2), 1-135.
4. Kim, Y. (2014). Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1746-1751.
5. Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent Convolutional Neural Networks for Text Classification. Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence.
6. Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2016). New Avenues in Opinion Mining and Sentiment Analysis. IEEE Intelligent Systems, 31(2), 76-81.
7. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics.
8. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. OpenAI Blog.