

Kinza dissetation_.docx

by Turnitin LLC

Submission date: 18-Nov-2024 10:33PM (UTC-0800)

Submission ID: 2524807197

File name: Kinza_dissetation_.docx (46.98K)

Word count: 3006

Character count: 17888

Title page

Project Title

Student Name:

Supervisor Name:

Acknowledgements

Abstract

Table of Contents

Introduction	8
Overview	8
Research Question	8
Aims and Objectives.....	9
1. Data Preparation	9
2. Model Development	9
3. Application Focus	9
Ethical Considerations.....	10
Background and Literature Review.....	11
Topic Overview	11
Review of Relevant Literature.....	11
Bibliography	16

Lists of Figures:

No table of figures entries found.

List of Tables:

Table 1 Litreature Reveiw Summary 14

Introduction

Overview

It has already been noticed that the Internet is growing at incredible speed, and it has caused the increase of the amount of text information continuously. Hosts like social networking sites, newsgroups, customer review sites, and email are creating huge volumes of text containing low-level or high-level feelings. Such feelings, whether expressed in words or only intelligible from context, are useful signals of the mental state, attitudes, or response to something. The interpretations and comprehension of these feelings have therefore emerged as paramount activities within the field of Natural Language Processing (NLP) (Chowdhary, 2020).

Speaking of emotion identification (Li, 2014), one needs to clarify that it concerns the identification of emotions within a text, which can be joy, anger, fear, sadness, etc. This task is different from general polarity summarization, which simply categorizes text as positive, negative, or neutral but tries to identify more relativistic emotions. For instance, though sentiment analysis may consider a statement such as “I am thrilled with the results” as positive, emotion classification would tag it simply as ‘joy.’ This level of classification differentiation enhances the creation of applications that can better suit the user needs and enhance decision-making systems.

It is crucial for emotion classification in a number of actual-world applications. In customer relationship management, for instance, analyzing passion in reviews or feedback enables companies to produce further relevant changes based on customer information. As it is the case with mental health analysis, the task of emotion classification can prove to be critical for the signs of depression, anxiety, or distress in written communication to be detected at the right time. Other potential areas of its application are human-computer interaction, where it will be useful to adjust responses to the mood of the participant in a dialogue, and education, where mood in feedback from the students can be helpful to adapt the learning approach.

However, the use of the emotion classification is not restricted as shown above because of the challenges posed by natural language processing. Emotions are personal, contextual, and may be multifaceted in a way, and this poses a challenge when it comes to identifying and labeling them. This is because of differences in writing technique, looseness in the use of language, regionalism and semantics, and lastly the use of proverbs, jokes, and riddles. The use of machine learning models in classifying emotions can be regarded as a potential solution to the stated challenges, as it can facilitate the automatic identification of emotions without compromising accuracy or extensibility (Bostan, 2018).

Research Question

The primary research question guiding this project is:

How can machine learning models be effectively applied to classify emotions in text data?

To address this question, the project explores the following sub-questions:

1. What are the most effective preprocessing techniques for handling textual data in emotion classification task?
2. How do different machine learning models, such as Naive Bayes, Support Vector Machines (SVM), and Random Forest, perform in emotion classification tasks?
3. What is the role of labeled datasets and feature engineering in enhancing classification accuracy?

Aims and Objectives

Hence, the main goal of this study is to create a machine learning framework for the classification of emotions based on textual data. The following specific objectives underpin this aim:

1. Data Preparation

The initial goal of this work is to select and acquire an adequate database for emotion recognition. The dataset should include text examples that are associated with emotions like “anger,” “joy,” or “fear.” Some of the operations required in preparing the dataset for model training include:

- First of all, preprocessing that additionally involves purging the text from stopwords, punctuation, and special characters.
- Interpreting the raw text into units of handling, such as words or phrases.
- Stemming or lemmatization is employed to transform words to their base form.
- Applying TLC with certain transformations, such as TF-IDF, to transform the text into a format that is comprehensible to ML.

These preprocessing steps help, in a way, to provide the data in the best condition possible so that models perform optimally.

2. Model Development

The second is to fine-tune deep learning models for performing the emotion classification for the text. This involves:

- Choosing simple methods as base models for initial testing, such as Naive Bayes, Random Forest, Adaboost, and finally SVM (Li X. L., 2008).
- Assessment of these models using parameters including accuracy, precision, recall, as well as F1-score.
- Adjusting the model settings to provide good results in the tasks dedicated to distinguishing emotions..

In this phase, it is also important to consider the further use of algorithmic strategies, including ensemble models and the hybrid method.

3. Application Focus

Since it concerns emotions, it has a broad impact on society, especially in the social, mental, and commercial platforms. This objective is to evaluate the realism of the models developed by:

- Using customer feedback to drive trends for better products and services.
- Identifying indicators of emotional experiencing in text-based communication to offer timely support in mental health settings.
- Improving people's communication with computers through making the technology sensitized to people's affective states.

It is through these applications that the project can guarantee that the results are not only scholarly but also useful.

Ethical Considerations

This project acknowledges that emotion classification is not without ethical concerns, especially when it comes to data privacy, bias, and consumption. The dataset employed in the study is also free from personally identifiable information and complies with GDPR. Precautions are made to prevent biases toward specific groups in the training data through using diversity to assess risks of misclassification for everyone.

To ensure models do not act in an opaque manner, the project focuses on the interpretability of the models, for instance, through feature importance analysis to explain classification. Ethical deployment is addressed around the same time, while some applications, such as mental health, have human oversight due to misclassification.

Through adherence to ethical AI principles, this work enforces responsible use of emotion classification technology, privacy, bias, and social good measures.

Background and Literature Review

Topic Overview

Sentiment analysis, a crucial component in the branch of NLP, is designed to identify and sort emotion signals within textual information. This shows that with the emergence of digital systems, emotions should be evaluated in large masses of information, such as social networks, customers' opinions, and messages. While traditional sentiment analysis classifies given text as positive, negative, or neutral, emotion classification tries to capture more subtle emotions like anger, joy, fear, or sadness. This specificity is used in such areas as the improvement of customer services, conducting a mental health check, and the interaction with a computer.

Emotion classification is a worthwhile yet difficult area of knowledge because it integrates both linguistic and psychological considerations. It is not easy to detect emotions because textual data is colored, context, and language change with context and often include nonstandard forms of language, such as slang and emoticons. Nevertheless, with the evolution of machine learning and deep learning, excellent tools and methods to solve these problems have been availed.

Review of Relevant Literature

This section critically discusses some of the major studies that have been conducted in the area of emotion classification and then relates them to the current study.

In this review, Acheampong et al. discuss the approaches and developments in emotion recognition from textual data with a focus on the issues and prospects for the field. The authors cover all the types of data useful for the emotion detection, including the datasets like ISEAR, SemEval datasets of 2007, 2017, and 2018, DailyDialog, etc. These datasets are famous for their emotional tag set, including features such as anger, joy, fear, sadness, and others. The fact that they are formatted differently from simple texts, like conversations to headlines, makes it possible to compare various methods used in emotion identification (Acheampong, 2020).

In this paper, the author reviews older machine-learning models such as SVM and Naive Bayes, which require a great deal of effort in feature engineering and optimal data structure. Such models tend to employ methods such as Term Frequency-Inverse Document Frequency (TF-IDF) and Bag-of-Words. But the review points out that CNNs and RNNs yield much better results than traditional techniques because of their capability to capture contextual and semantic information. For example, BiLSTM models coupled with attention yielded high performance, reaching 91% accuracy in structured datasets such as SemEval. However, standard approaches in ML generate moderate accuracy of between 70% and 80%, especially when dealing with divergent or unorganized data.

The paper also points out several limitations, such as lack of data, ambiguity of languages, and identification of mixed signals. To overcome such limitations, Acheampong et al. propose using hybrid configurations of lexicon-based techniques coupled with machine learning or deep learning strategies. These mixed models can take advantage of better interpretability of lexicon-based

methods as well as better flexibility and contextual understanding of neural networks. The findings of this paper can be applied directly to the present work, focusing on dataset choice and the hybrid models for emotion categorization (Acheampong, 2020).

Murthy and Kumar's review categorizes emotion detection methodologies into three primary approaches: lexicon-based methods, traditional ML models, and Deep Learning models. In the paper, the authors described several datasets such as CrowdFlower, SemEval, and DailyDialog, using which textual data is annotated with different emotions. These datasets are diverse and include formal spoken language, tweets, and casual conversation, which is useful for comparing models in various settings (Murthy, 2021).

Interpretable and simple, though lexicon-based approaches are marked for their poor scalability and incapability to capture context. For instance, such methods base their analysis on particular dictionaries that might not include new words, phrases, slang, and idioms, among other things. On the other hand, there are approaches like SVM and Random Forest that use techniques from feature engineering even with the use of TF-IDF and Word2Vec 2Vec for records. However, their reliance on tabular-formatted data and usually requiring significant feature engineering reduce their usefulness with text data.

The authors also discuss how deep learning models, including CNNs and LSTMs, help to overcome these limitations. These models are said to gain accuracy on a scale of 85%–95% on corpora like SemEval through the use of pre-trained embeddings like GloVe as well as contextualized ones like BERT. This research also discusses efficiency when combining lexicon-based methods with neural networks, which offer optimal accuracy with manageable interpretability issues. These insights are useful for the current project, as they emphasize the need to integrate conventional approaches and deep learning methods to obtain accurate emotion classification (Murthy, 2021).

Hogenboom et al. expand on the use of emoticons that contribute to the improvement of the semantic analysis. The study utilizes two distinct datasets: 2,080 Dutch tweets and forum messages labeled with emoticons and 10,069 English app reviews, some of which contain emoticons as text-based sentiment markers. These datasets were selected with the intent of understanding the effect of emoticons on sentiment as well as emotional analysis (Hogenboom, 2015).

The authors recommend a lexical resource supported by an emoticon sentiment lexicon, which is then combined with basic bag-of-words approaches. This fusion technique enhances polarity identification and provides the results, turning from 68% to 72% in bipolar classification. The addition of emoticon sentiment improves the performance of the system by 10–15% over systems that do not consider emoticons. This study underlines the importance of non-verbal communication, such as emoticons, when using written data, as they reflect intensity for the textual content.

This research is relevant in understanding the implication of multimodal data in emotion recognition. Its specialization in exploring textual as well as emoticon features is parallel to the current project's objective of utilizing both the verbal and non-verbal indicators of the emotion.

Thus, the work can use combinatorial approaches to describe the contextual variations of the textual data, which occur in social networks and other informal environments (Hogenboom, 2015).

Batbaatar et al. propose the Semantic-Emotion Neural Network (SENN), a double-path model for overcoming the challenges that single-path models present when it comes to emotion classification. The study employs standard datasets such as SemEval, annotated with Ekman's six basic emotions: pleasure, interest, anger, hurt, suspense, and loathsome. These datasets are effective in measuring the effectiveness of the various approaches to emotion classification (Batbaatar, 2019).

SENN integrates two sub-networks:

1. **BiLSTM (Bidirectional Long Short-Term Memory):** Captures semantic contextual relationships within text.
2. **CNN (Convolutional Neural Network):** Extracts position-invariant features, focusing on emotional patterns.

The model employs word vectors like Word2Vec, GloVe, and FastText that are retrained for better results. The results of the compared experiments show that SENN is superior to other models, including Naive Bayes and SVM, from the perspective of precision, recall, and F1-score. It is also capable of both semantic and emotional features, so it provides a fair shot at managing short and noisy text.

The connection of this study lies in the fact that it develops a dual-network architecture that can help in semantic and emotional feature integration. In the same vein, the project could improve upon the techniques of emotion classification by adopting such means of learning (Batbaatar, 2019).

3 This paper proposes a new idea of sentiment classification based on combining emoticons and short text content. The authors collect a data set of 100,000 tweets from Twitter, then clean the data by deleting links and usernames and stripping away short tweets with less than 3 words. The last dataset consists of 2000 positive and 2000 negative tweets containing at least one emoji (Zou, 2022).

The proposed methodology involves:

1. **Emoji Vectorization Algorithm:** Adapts Word2Vec to train emoji-specific vectors, capturing semantic and emotional information.
2. **Convolutional Neural Networks (CNNs):** Combines text and emoji vectors into sentence matrices for enhanced feature extraction.

The CNN model achieves a 10–12% accuracy improvement over baseline methods such as Naive Bayes and basic word embedding approaches. The study effectively addresses limitations in sentiment analysis, such as data sparsity and the lack of context in short texts. By incorporating emoticons, it enhances emotional understanding, making it particularly relevant for social media analysis.

This paper's emphasis on blending multimodal data provides a valuable framework for integrating textual and non-textual features in emotion classification, aligning closely with the objectives of the current project (Zou, 2022).

Here is the tabular representation of the literature review summarizing the key points from each paper:

Table 1 Literature Review Summary

Paper	Dataset	Models/Techniques	Results	Relevance to Project
(Acheampong, 2020)	ISEAR, SemEval (2007, 2017, 2018), and DailyDialog datasets, annotated with emotions like anger, joy, fear, and sadness.	SVM, Naive Bayes, CNN, Bi-LSTM with attention mechanisms, feature extraction using TF-IDF and Bag-of-Words.	Bi-LSTM with attention achieved >90% accuracy on structured datasets (e.g., SemEval); traditional models achieved 70–80%.	Highlights challenges like data diversity and proposes hybrid approaches combining lexicon-based and deep learning techniques, aligning with project objectives.
(Murthy, 2021)	CrowdFlower, SemEval, and DailyDialog datasets, covering diverse contexts such as social media and formal conversations.	Lexicon-based methods, traditional models (SVM, Random Forest), deep learning (CNN, LSTM), embeddings (GloVe, BERT).	Deep learning models with pre-trained embeddings (e.g., GloVe, BERT) achieved 85–95% accuracy; lexicon-based methods performed at 60–70% accuracy.	Advocates hybrid approaches that balance interpretability and performance, emphasizing deep learning with embeddings for robust emotion classification.
(Hogenboom, 2015)	2,080 Dutch tweets and forum messages, 10,069 English app reviews annotated with emoticons and sentiment polarity.	Lexicon-based approach enhanced by an emoticon sentiment lexicon combined with bag-of-words.	Achieved 68–72% accuracy for binary classification; incorporating emoticons improved accuracy by 10–15% compared to text-only models.	Demonstrates the importance of leveraging non-verbal cues like emoticons, supporting multimodal approaches in emotion detection for real-world social media data.
(Batbaatar, 2019)	Standard datasets (e.g., SemEval) annotated with Ekman's six basic emotions: joy, fear, anger, sadness, surprise, disgust.	Semantic-Emotion Neural Network (SENN) combining BiLSTM for contextual relationships and CNN for emotional patterns.	SENN outperformed Naive Bayes and SVM in terms of precision, recall, and F1-score; robust integration of semantic and emotional features in noisy data.	Provides a dual-network framework integrating semantic and emotional cues, aligning with project goals of robust feature extraction and classification accuracy.
(Zou, 2022)	100,000 tweets processed to include 2,000 positive and	Emoji Vectorization Algorithm (adapts Word2Vec for emojis) and CNN	CNN achieved 10–12% accuracy improvement over baseline methods	Supports multimodal approaches blending text and emojis, aligning with the

	2,000 negative tweets annotated with emojis.	combining text and emoji vectors into sentence matrices.	(Naive Bayes, traditional word embeddings); addressed data sparsity and context in short texts.	project's focus on leveraging both textual and non-textual features in emotion analysis.
--	--	--	---	--

This table summarizes the datasets, methodologies, results, and relevance of each paper in an accessible and structured format for easy comparison and application to your project.

Bibliography

- ¹³ Chowdhary, K. a. (2020). Natural language processing. *Fundamentals of artificial intelligence* , 603-649.
- ⁶ Li, W. a. (2014). Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications* 41.4 (2014): , 1742-1749.
- ¹⁵ Bostan, L. A. (2018). An analysis of annotated corpora for emotion classification in text. *Bamberg: Otto-Friedrich-Universität*.
- ⁷ Li, X. L. (2008). AdaBoost with SVM-based component classifiers. *Engineering Applications of Artificial Intelligence* 21.5 , 785-795.
- ² Acheampong, F. A.-M. (2020). Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports* 2.7 , e12189.
- Murthy, A. R. (2021). A review of different approaches for detecting emotion from text. *IOP Conference Series: Materials Science and Engineering. Vol. 1110. No. 1. IOP Publishing* .
- ⁵ Hogenboom, A. D. (2015). Exploiting emoticons in polarity classification of text. *Journal of Web Engineering* , 022-040.
- ⁸ Batbaatar, E. M. (2019). Semantic-emotion neural network for emotion recognition from text. *IEEE access* 7, 111866-111878.
- ⁹ Zou, H. a. (2022). Sentiment classification method based on blending of emoticons and short texts. *Entropy* 24.3 , 398.

Kinza dissetation_.docx

ORIGINALITY REPORT

10%

SIMILARITY INDEX

10%

INTERNET SOURCES

7%

PUBLICATIONS

6%

STUDENT PAPERS

PRIMARY SOURCES

1

louisdl.louislibraries.org

Internet Source

1%

2

link.springer.com

Internet Source

1%

3

www.mdpi.com

Internet Source

1%

4

ijeecs.iaescore.com

Internet Source

1%

5

pure-accept.eur.nl

Internet Source

1%

6

www.researchgate.net

Internet Source

1%

7

hal-ifp.archives-ouvertes.fr

Internet Source

1%

8

Submitted to University of Bedfordshire

Student Paper

1%

9

d197for5662m48.cloudfront.net

Internet Source

1%

10	fastercapital.com Internet Source	1 %
11	iieta.org Internet Source	1 %
12	www.coursehero.com Internet Source	<1 %
13	ijcaonline.org Internet Source	<1 %
14	C. Akalya devi, D. Karthika Renuka, G. Pooventhiran, D. Harish, Shweta Yadav, Krishnaprasad Thirunarayan. "Towards enhancing emotion recognition via multimodal framework", Journal of Intelligent & Fuzzy Systems, 2023 Publication	<1 %
15	deepai.org Internet Source	<1 %
16	perform.concordia.ca Internet Source	<1 %
17	ir.psych.ac.cn Internet Source	<1 %
18	aclanthology.org Internet Source	<1 %
19	listens.online Internet Source	<1 %

Exclude quotes Off

Exclude matches Off

Exclude bibliography Off