Sentiment Analysis of Hotel Reviews Using Deep Learning

Nathan Baxter  
*School of Computing*   
*Ulster University   
Belfast, United Kingdom*baxter-n3@ulster.ac.uk

*Abstract*—This paper presents a sentiment classification system for hotel reviews using deep learning approaches. The system analyzes TripAdvisor hotel reviews and classifies them into positive, negative, and neutral sentiment categories. Two deep neural network architectures—Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM)—are implemented and compared to determine the most effective approach. The preprocessing pipeline includes text cleaning, stopword removal, lemmatization, and word embedding. The models are evaluated using a 60%-40% train-test split, with performance metrics including accuracy, precision, recall, and F1-score. The LSTM model achieved slightly better performance with an accuracy of 82.81% compared to BiLSTM's 82.77%. Analysis of the confusion matrix reveals strong performance for positive reviews (94% recall) but challenges with neutral sentiment detection. This paper details the implementation process, compares model performance, and discusses approaches to address class imbalance issues in sentiment classification tasks.

Keywords-sentiment analysis, deep learning, LSTM, BiLSTM, hotel reviews, natural language processing, TensorFlow, Keras.

# Introduction

Sentiment analysis has become an essential tool for businesses to understand customer opinions and experiences expressed in online reviews. The hospitality industry, in particular, relies heavily on customer feedback to improve services and maintain competitive advantage. Traditional approaches to sentiment analysis often involve manual inspection or rule-based systems, which become impractical with large volumes of data.

This paper focuses on the implementation of deep learning models for sentiment classification of hotel reviews from TripAdvisor. The objective is to classify reviews into three sentiment categories: positive, negative, and neutral. We implement and compare two deep neural network architectures: Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) networks.

The dataset used in this study consists of 20,491 TripAdvisor hotel reviews, each with a corresponding rating (1-5). The preprocessing pipeline includes text cleaning, stopword removal, lemmatization, and tokenization. The processed text is then converted into numerical representations using word embeddings.

By comparing the performance of LSTM and BiLSTM models, we aim to identify the most effective approach for hotel review sentiment classification. The evaluation includes accuracy, precision, recall, and F1-score metrics, as well as analysis of the confusion matrix to understand misclassification patterns.

# SENTIMENT CLASSIFICATION PROCESS

## Dataset Overview

The TripAdvisor hotel reviews dataset used in this study contains 20,491 reviews with corresponding ratings on a 1-5 scale. The dataset has two columns: "Review" (text content) and "Rating" (integer rating). The ratings were converted to sentiment labels as follows:

- Ratings 1-2: Negative sentiment (15.7% of reviews)

- Rating 3: Neutral sentiment (10.7% of reviews)

- Ratings 4-5: Positive sentiment (73.7% of reviews)

This distribution reveals a significant class imbalance favouring positive reviews, which is common in hotel review datasets but presents challenges for model training.

## Preprocessing Pipeline

The preprocessing pipeline includes several steps to transform raw text data into a format suitable for deep learning models:

1) Text Cleaning: This step removes special characters, HTML tags, URLs, and numbers from the review text. All text is converted to lowercase, and extra whitespace is removed. This standardization helps reduce the vocabulary size and focuses on meaningful content.

```python

def clean\_text(text):

if not isinstance(text, str):

return ""

# Convert to lowercase

text = text.lower()

# Remove HTML tags if any

text = re.sub(r'', '', text)

# Remove URLs

text = re.sub(r'http\S+|www\S+|https\S+', '', text)

# Remove special characters and numbers

text = re.sub(r'[^a-zA-Z\s]', '', text)

# Remove extra whitespace

text = re.sub(r'\s+', ' ', text).strip()

return text

```

2) Stopword Removal: Common words (e.g., "the", "a", "an") that do not contribute significantly to sentiment are removed using NLTK's stopwords list.

```python

def remove\_stopwords(text):

if not isinstance(text, str) or text == "":

return ""

stop\_words = set(stopwords.words('english'))

word\_tokens = text.split()

filtered\_text = [word for word in word\_tokens if word not in stop\_words]

return ' '.join(filtered\_text)

```

3) Lemmatization: Words are reduced to their base or dictionary form using NLTK's WordNetLemmatizer. This helps standardize different forms of the same word (e.g., "staying", "stayed", "stays" all become "stay").

```python

def lemmatize\_text(text):

if not isinstance(text, str) or text == "":

return ""

lemmatizer = WordNetLemmatizer()

word\_tokens = text.split()

lemmatized\_text = [lemmatizer.lemmatize(word) for word in word\_tokens]

return ' '.join(lemmatized\_text)

```

4) Tokenization and Padding: The preprocessed text is converted into sequences of numerical tokens using Keras Tokenizer. These sequences are then padded to ensure uniform length (100 tokens) for model input.

```python

def tokenize\_and\_pad(texts, max\_words=10000, max\_sequence\_length=100):

tokenizer = Tokenizer(num\_words=max\_words)

tokenizer.fit\_on\_texts(texts)

sequences = tokenizer.texts\_to\_sequences(texts)

padded\_sequences = pad\_sequences(sequences, maxlen=max\_sequence\_length)

return tokenizer, padded\_sequences

```

## Word Embedding

Word embedding is a critical feature representation technique that maps words to dense vectors in a high-dimensional space. The embedding layer in our neural network models converts word indices to dense vectors of fixed size (128 dimensions). This representation captures semantic relationships between words, where words with similar meanings have similar vector representations.

Rather than using pre-trained embeddings like GloVe or Word2Vec, our implementation learns word embeddings during model training, allowing the representations to be specifically tailored to the hotel review domain. The embedding layer is the first layer in both the LSTM and BiLSTM models:

```python

model.add(Embedding(max\_features, embedding\_dim, input\_length=max\_len))

model.add(SpatialDropout1D(0.2))

```

Where `max\_features` is the vocabulary size (70,983 words in our implementation), `embedding\_dim` is set to 128, and `max\_len` is the maximum sequence length (100).

# MODEL ARCHITECTURE AND SELECTION

## LSTM Model Architecture

The Long Short-Term Memory (LSTM) model is designed to capture sequential dependencies in text. The architecture includes:

1) Embedding Layer: Converts word indices to dense vectors

2) SpatialDropout1D: Reduces overfitting by randomly dropping word embeddings

3) LSTM Layer: Processes sequential data with 128 units

4) Dense Layer: 64 neurons with ReLU activation

5) Dropout Layer: 50% dropout for regularization

6) Output Layer: 3 neurons with softmax activation for classification

```python

def create\_lstm\_model(max\_features, embedding\_dim=128, max\_len=100, num\_classes=3):

model = Sequential()

model.add(Embedding(max\_features, embedding\_dim, input\_length=max\_len))

model.add(SpatialDropout1D(0.2))

model.add(LSTM(128, dropout=0.2, recurrent\_dropout=0.2))

model.add(Dense(64, activation='relu'))

model.add(Dropout(0.5))

model.add(Dense(num\_classes, activation='softmax'))

model.compile(loss='categorical\_crossentropy', optimizer='adam', metrics=['accuracy'])

return model

```

## BiLSTM Model Architecture

* Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
* Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
* Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter”, not “webers/m2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.

Identify applicable funding agency here. If none, delete this text box.

* Use a zero before decimal points: “0.25”, not “.25”. Use “cm3”, not “cc”. (*bullet list*)

## Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

*a**b* 

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

## Some Common Mistakes

* The word “data” is plural, not singular.
* The subscript for the permeability of vacuum **0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
* In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
* A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
* Do not use the word “essentially” to mean “approximately” or “effectively”.
* In your paper title, if the words “that uses” can accurately replace the word “using”, capitalize the “u”; if not, keep using lower-cased.
* Be aware of the different meanings of the homophones “affect” and “effect”, “complement” and “compliment”, “discreet” and “discrete”, “principal” and “principle”.
* Do not confuse “imply” and “infer”.
* The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
* There is no period after the “et” in the Latin abbreviation “et al.”.
* The abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

An excellent style manual for science writers is [7].

# Using the Template

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Authors and Affiliations

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

### For papers with less than six authors: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

#### Deletion: Delete the author and affiliation lines for the extra authors.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. *(references)*
2. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
3. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.
4. K. Elissa, “Title of paper if known,” unpublished.
5. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
6. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].
7. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi TIFF or EPS file, with all fonts embedded) because, in an MSW document, this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord “Format” pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.