Employing Transformers and Emoji to Perform Sentiment Classification of Social Media Texts

Master's Presentation

Tiago Martinho de Barros
Supervisor: Prof. Dr. Hélio Pedrini
Co-supervisor: Prof. Dr. Zanoni Dias

University of Campinas
Institute of Computing

May 3, 2021

Agenda

Agenda


Mark Duffy
@LordduFey
Replying to @F1
Schumacher was such a dirty driver.
24 7:01 PM - Jul 28, 2018


Mark Duffy
@LordduFey
Replying to @F1
Schumacher was such a dirty driver.
24 7:01 PM - Jul 28, 2018

Introduction

Examples

Introduction

Examples

- Sentiment Analysis.
  - Positive, neutral, or negative.
- Text types:
  - Objective: does not contain opinion.
    E.g.: "Winter is coming".
  - Subjective: does contain opinion.
    E.g.: "Such a magnificent kimono she is wearing".
- 216 million messages are shared by Facebook users[1] and 500 million tweets are sent[2] every day.

---

[1] Statista, *User-generated Internet content per minute as of August 2020*, 2020, Accessed: 2021-04-26.

[2] Internet Live Stats, *Twitter Usage Statistics*, 2021, Accessed: 2021-04-26.

Introduction

Problem Description

- Sentiment Analysis.
  - Positive, neutral, or negative.
- Text types:
  - Objective: does not contain opinion.
    E.g.: “Winter is coming.” ❄️
  - Subjective: does contain opinion.
    E.g.: “Such a magnificent kimono she is wearing.”
- 216 million messages are shared by Facebook users[1] and 500 million tweets are sent[2] every day.

---

[1] Statista, *User-generated Internet content per minute as of August 2020*, 2020, Accessed: 2021-04-26.

[2] Internet Live Stats, *Twitter Usage Statistics*, 2021, Accessed: 2021-04-26.

Introduction

Problem Description

- Sentiment Analysis.
  - Positive, neutral, or negative.
- Text types:
  - Objective: does not contain opinion.
    E.g.: "Winter is coming." ❄️
  - Subjective: does contain opinion.
    E.g.: "Such a magnificent kimono she is wearing." 👘
- 216 million messages are shared by Facebook users[1] and 500 million tweets are sent[2] every day.

---

[1] Statista, *User-generated Internet content per minute as of August 2020*, 2020, Accessed: 2021-04-26.

[2] Internet Live Stats, *Twitter Usage Statistics*, 2021, Accessed: 2021-04-26.

Introduction

Problem Description

- Sentiment Analysis.
  - Positive, neutral, or negative.
- Text types:
  - Objective: does not contain opinion.
    E.g.: "Winter is coming." ❄️
  - Subjective: does contain opinion.
    E.g.: "Such a magnificent kimono she is wearing." 👘
- 216 million messages are shared by Facebook users[1] and 500 million tweets are sent[2] every day.

[1] Statista, *User-generated Internet content per minute as of August 2020*, 2020, Accessed: 2021-04-26.

[2] Internet Live Stats, *Twitter Usage Statistics*, 2021, Accessed: 2021-04-26.

Introduction

Motivation

- By 2025, 463 exabytes (EB) of data will be created every day.[1]
- Examples of practical uses: gauging user opinion about products and services, building recommender systems[2] and predicting the outcome of an election.[3,4]
- Shifts in sentiment on social media have been shown to correlate with shifts in the stock market.[5]

---

[1] Raconteur, *A Day in Data*, 2020, Accessed: 2021-04-26.

[2] Selmene and Kodia, "Recommender System Based on User's Tweets Sentiment Analysis", In *4th International Conference on E-Commerce, E-Business and E-Government (ICEEG)*, 2020.

[3] Cristiani, Lieira, and Camargo, "A Sentiment Analysis of Brazilian Elections Tweets", In *Proceedings of the 8th Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)*, 2020.

[4] Liu, Yao, Guo, and Wei, "Can We Forecast Presidential Election Using Twitter Data? An Integrative Modelling Approach", *Annals of GIS*, 2020.

[5] Ingle, Kante, Samak, and Kumari, "Sentiment Analysis of Twitter Data Using Hadoop", *International Journal of Engineering Research and General Science*, 2015.

Introduction

Motivation

- By 2025, 463 exabytes (EB) of data will be created every day.[1]
- Examples of practical uses: gauging user opinion about products and services, building recommender systems[2] and predicting the outcome of an election.[3,4]
- Shifts in sentiment on social media have been shown to correlate with shifts in the stock market.[5]

[1] Raconteur, *A Day in Data*, 2020, Accessed: 2021-04-26.

[2] Selmene and Kodia, "Recommender System Based on User's Tweets Sentiment Analysis", In *4th International Conference on E-Commerce, E-Business and E-Government (ICEEG)*, 2020.

[3] Cristiani, Lieira, and Camargo, "A Sentiment Analysis of Brazilian Elections Tweets", In *Proceedings of the 8th Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)*, 2020.

[4] Liu, Yao, Guo, and Wei, "Can We Forecast Presidential Election Using Twitter Data? An Integrative Modelling Approach", *Annals of GIS*, 2020.

[5] Ingle, Kante, Samak, and Kumari, "Sentiment Analysis of Twitter Data Using Hadoop", *International Journal of Engineering Research and General Science*, 2015.

Introduction

- By 2025, 463 exabytes (EB) of data will be created every day.[1]
- Examples of practical uses: gauging user opinion about products and services, building recommender systems[2] and predicting the outcome of an election.[3,4]
- Shifts in sentiment on social media have been shown to correlate with shifts in the stock market.[5]

😮

[1] Raconteur, *A Day in Data*, 2020, Accessed: 2021-04-26.

[2] Selmene and Kodia, "Recommender System Based on User's Tweets Sentiment Analysis", In *4th International Conference on E-Commerce, E-Business and E-Government (ICEEG)*, 2020.

[3] Cristiani, Lieira, and Camargo, "A Sentiment Analysis of Brazilian Elections Tweets", In *Proceedings of the 8th Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)*, 2020.

[4] Liu, Yao, Guo, and Wei, "Can We Forecast Presidential Election Using Twitter Data? An Integrative Modelling Approach", *Annals of GIS*, 2020.

[5] Ingle, Kante, Samak, and Kumari, "Sentiment Analysis of Twitter Data Using Hadoop", *International Journal of Engineering Research and General Science*, 2015.

Introduction

Objectives

- **General objective**: Investigate state-of-the-art Sentiment Analysis techniques, and contribute to the field.
- **Specific objectives**:
  - Search for suitable datasets.
  - Evaluation of recent works on sentiment classification and on NLP that can be used to perform sentiment classification.
  - Proposition of an original methodology to classify the sentiment of social media texts.
  - Conduction of experiments on data augmentation.
  - Performance evaluation of the developed model.

- **General objective**: Investigate state-of-the-art Sentiment Analysis techniques, and contribute to the field.
- **Specific objectives**:
  - Search for suitable datasets.
  - Evaluation of recent works on sentiment classification and on NLP that can be used to perform sentiment classification.
  - Proposition of an original methodology to classify the sentiment of social media texts.
  - Conduction of experiments on data augmentation.
  - Performance evaluation of the developed model.

- **General objective**: Investigate state-of-the-art Sentiment Analysis techniques, and contribute to the field.
- **Specific objectives**:
  - Search for suitable datasets.
  - Evaluation of recent works on sentiment classification and on NLP that can be used to perform sentiment classification.
  - Proposition of an original methodology to classify the sentiment of social media texts.
  - Conduction of experiments on data augmentation.
  - Performance evaluation of the developed model.

- **General objective**: Investigate state-of-the-art Sentiment Analysis techniques, and contribute to the field.
- **Specific objectives**:
  - Search for suitable datasets.
  - Evaluation of recent works on sentiment classification and on NLP that can be used to perform sentiment classification.
  - Proposition of an original methodology to classify the sentiment of social media texts.
  - Conduction of experiments on data augmentation.
  - Performance evaluation of the developed model.

Introduction

Objectives

- **General objective**: Investigate state-of-the-art Sentiment Analysis techniques, and contribute to the field.
- **Specific objectives**:
  - Search for suitable datasets.
  - Evaluation of recent works on sentiment classification and on NLP that can be used to perform sentiment classification.
  - Proposition of an original methodology to classify the sentiment of social media texts.
  - Conduction of experiments on data augmentation.
  - Performance evaluation of the developed model.

Introduction

Objectives

- **General objective**: Investigate state-of-the-art Sentiment Analysis techniques, and contribute to the field.
- **Specific objectives**:
    - Search for suitable datasets.
    - Evaluation of recent works on sentiment classification and on NLP that can be used to perform sentiment classification.
    - Proposition of an original methodology to classify the sentiment of social media texts.
    - Conduction of experiments on data augmentation.
    - Performance evaluation of the developed model.

1. Do emoji, considered alongside their corresponding texts, improve the sentiment classification accuracy?
2. Can further unsupervised pre-training on in-domain data improve the sentiment classification performance?
3. Considering that unsupervised language representation learning methods are pre-trained on gigabytes of textual data, does data augmentation improve the sentiment classification performance?

Introduction

Research Questions

1. Do emoji, considered alongside their corresponding texts, improve the sentiment classification accuracy?
2. Can further unsupervised pre-training on in-domain data improve the sentiment classification performance?
3. Considering that unsupervised language representation learning methods are pre-trained on gigabytes of textual data, does data augmentation improve the sentiment classification performance?

Introduction

Research Questions

1. Do emoji, considered alongside their corresponding texts, improve the sentiment classification accuracy?
2. Can further unsupervised pre-training on in-domain data improve the sentiment classification performance?
3. Considering that unsupervised language representation learning methods are pre-trained on gigabytes of textual data, does data augmentation improve the sentiment classification performance?

🤔

- A study of emoji occurrence and distribution for the most frequent emoji in the TweetSentBR and the 2000-tweets-BR datasets, comparing the results with general emoji usage in Twitter.
- A novel methodology to classify the sentiment of social media texts using both the expressiveness of emoji and the written text. Our model achieves a new state of the art for both datasets.
- Despite being a different model, we can reduce the training time by using a previously pre-trained $\text{BERT}_{\text{BASE}}$ model to warm start ours, thus avoiding having to pre-train it from scratch.

Introduction

Contributions

- A study of emoji occurrence and distribution for the most frequent emoji in the TweetSentBR and the 2000-tweets-BR datasets, comparing the results with general emoji usage in Twitter.
- A novel methodology to classify the sentiment of social media texts using both the expressiveness of emoji and the written text. Our model achieves a new state of the art for both datasets.
- Despite being a different model, we can reduce the training time by using a previously pre-trained $\text{BERT}_{\text{BASE}}$ model to warm start ours, thus avoiding having to pre-train it from scratch.

- A study of emoji occurrence and distribution for the most frequent emoji in the TweetSentBR and the 2000-tweets-BR datasets, comparing the results with general emoji usage in Twitter.
- A novel methodology to classify the sentiment of social media texts using both the expressiveness of emoji and the written text. Our model achieves a new state of the art for both datasets.
- Despite being a different model, we can reduce the training time by using a previously pre-trained $\text{BERT}_{\text{BASE}}$ model to warm start ours, thus avoiding having to pre-train it from scratch.

Agenda

- Document-level sentiment classification.
- Sentence-level sentiment classification.
- Aspect-level sentiment classification.
  - *"The art direction of 'Star Wars: The Force Awakens' was amazing, but the plot was uninteresting, to say the least."*

Background

Sentiment Analysis Scope

- Document-level sentiment classification.
- Sentence-level sentiment classification.
- Aspect-level sentiment classification.
  - *"The art direction of 'Star Wars: The Force Awakens' was amazing, but the plot was uninteresting, to say the least."*

- Document-level sentiment classification.
- Sentence-level sentiment classification.
- Aspect-level sentiment classification.
  - *"The art direction of 'Star Wars: The Force Awakens' was amazing, but the plot was uninteresting, to say the least."*

Background

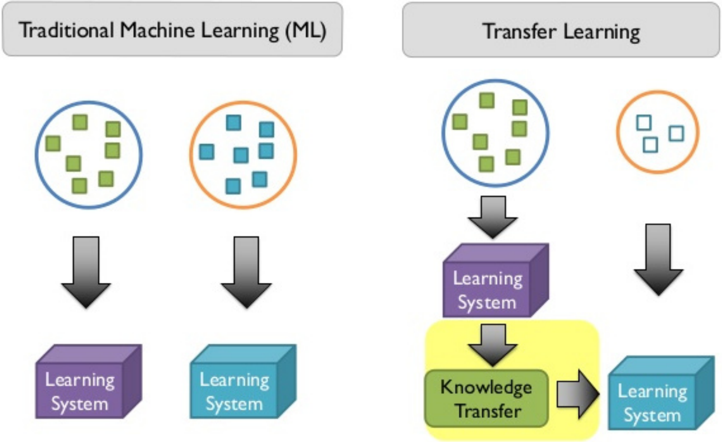
Sentiment Analysis Approaches

- Rule-based approach.
    - E.g.: if it contains “great” $\Rightarrow$ positive sentiment.
- Lexicon-based approach.
    - E.g.: $\begin{array}{ccc} \text{She} & \text{is} & \text{nice} \\ 0 & 0 & +3 \end{array} \Rightarrow 3 > 0 \Rightarrow$ positive sentiment.
- Machine Learning-based approach.
    - Classical Machine Learning approach.
    - Deep Learning approach.

- Rule-based approach.
  - E.g.: if it contains “great” $\Rightarrow$ positive sentiment.
- Lexicon-based approach.
  - E.g.: $\begin{array}{ccc} \text{She} & \text{is} & \text{nice.} \\ 0 & 0 & +3 \end{array} \Rightarrow 3 > 0 \Rightarrow$ positive sentiment.
- Machine Learning-based approach.
  - Classical Machine Learning approach.
  - Deep Learning approach.

- Rule-based approach.
  - E.g.: if it contains “great” $\Rightarrow$ positive sentiment.
- Lexicon-based approach.
  - E.g.: $\begin{array}{ccc} \text{She} & \text{is} & \text{nice.} \\ 0 & 0 & +3 \end{array} \Rightarrow 3 > 0 \Rightarrow$ positive sentiment.
- Machine Learning-based approach.
  - Classical Machine Learning approach.
  - Deep Learning approach.

- Rule-based approach.
  - E.g.: if it contains "great" $\Rightarrow$ positive sentiment.
- Lexicon-based approach.
  - E.g.: $\begin{array}{ccc} \text{She} & \text{is} & \text{nice.} \\ 0 & 0 & +3 \end{array} \Rightarrow 3 > 0 \Rightarrow$ positive sentiment.
- Machine Learning-based approach.
  - Classical Machine Learning approach.
  - Deep Learning approach.



An illustration of Transfer Learning.[1]

[1] Anchit Jain, "Improve Your Model Accuracy by Transfer Learning", 2018, Accessed: 2019-09-27.

An illustration of Language Modeling.[1]

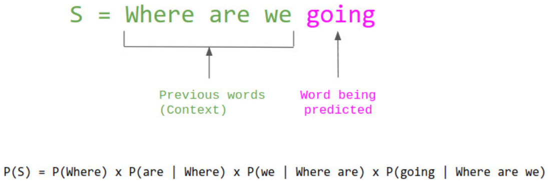
[1] Munroe, *What If? #34 Twitter*, 2013, Accessed: 2021-04-22.

[2] Huyen, *Evaluation Metrics for Language Modeling*, 2019, Accessed: 2021-04-22.

Background

Language Modeling

An illustration of Language Modeling.[1]



Example of Language Modeling.[2]

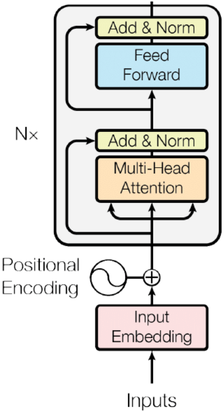
[1]Munroe, *What If? #34 Twitter*, 2013, Accessed: 2021-04-22.

[2]Huyen, *Evaluation Metrics for Language Modeling*, 2019, Accessed: 2021-04-22.

Background

Transformer

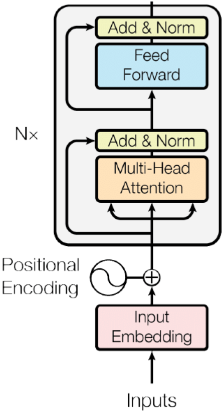
- Transformer[1] is a sequence transduction model.
- Examples: Speech Recognition, Text-To-Speech, and Machine Translation.
- Originally developed to perform Neural Machine Translation.
- It relies entirely on self-attention to compute representations of its input and output instead of recurrence.
- Thus, it lends itself better to parallelization than Recurrent Neural Networks.



Transformer encoder.

[1] Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin, "Attention is All You Need", In *31st Conference on Neural Information Processing Systems (NIPS)*, 2017.

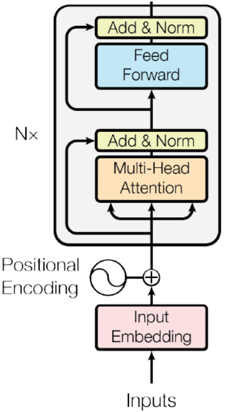
- Transformer[1] is a sequence transduction model.
- Examples: Speech Recognition, Text-To-Speech, and Machine Translation.
- Originally developed to perform Neural Machine Translation.
- It relies entirely on self-attention to compute representations of its input and output instead of recurrence.
- Thus, it lends itself better to parallelization than Recurrent Neural Networks.



Transformer encoder.

[1] Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin, "Attention is All You Need", In *31st Conference on Neural Information Processing Systems (NIPS)*, 2017.

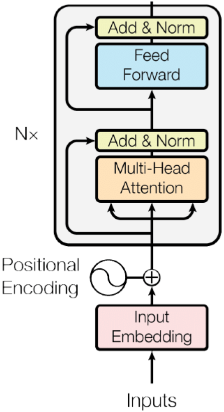
- Transformer[1] is a sequence transduction model.
- Examples: Speech Recognition, Text-To-Speech, and Machine Translation.
- Originally developed to perform Neural Machine Translation.
- It relies entirely on self-attention to compute representations of its input and output instead of recurrence.
- Thus, it lends itself better to parallelization than Recurrent Neural Networks.



Transformer encoder.

[1]Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin, "Attention is All You Need", In *31st Conference on Neural Information Processing Systems (NIPS)*, 2017.

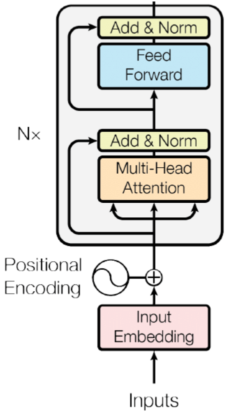
- Transformer[1] is a sequence transduction model.
- Examples: Speech Recognition, Text-To-Speech, and Machine Translation.
- Originally developed to perform Neural Machine Translation.
- It relies entirely on self-attention to compute representations of its input and output instead of recurrence.
- Thus, it lends itself better to parallelization than Recurrent Neural Networks.



Transformer encoder.

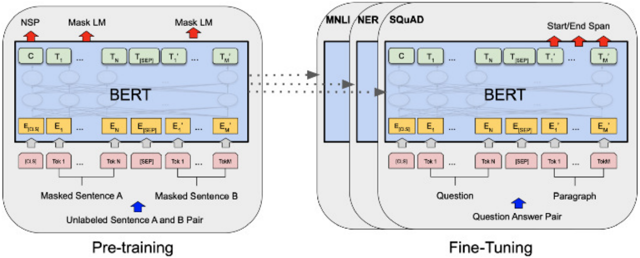
[1] Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin, "Attention is All You Need", In *31st Conference on Neural Information Processing Systems (NIPS)*, 2017.

- Transformer[1] is a sequence transduction model.
- Examples: Speech Recognition, Text-To-Speech, and Machine Translation.
- Originally developed to perform Neural Machine Translation.
- It relies entirely on self-attention to compute representations of its input and output instead of recurrence.
- Thus, it lends itself better to parallelization than Recurrent Neural Networks.



Transformer encoder.

[1] Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin, "Attention is All You Need", In *31st Conference on Neural Information Processing Systems (NIPS)*, 2017.



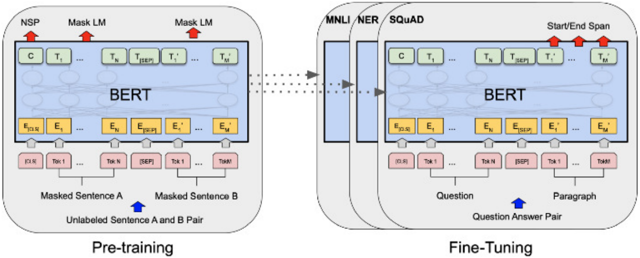
An illustration of Bidirectional Encoder Representations from Transformers (BERT).

- BERT[1] employs a bidirectional Transformer.
- Pre-trained using two unsupervised tasks:

[1] Devlin, Chang, Lee, and Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", In *20th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT): Long and Short Papers - Volume 1*, 2019.

Background

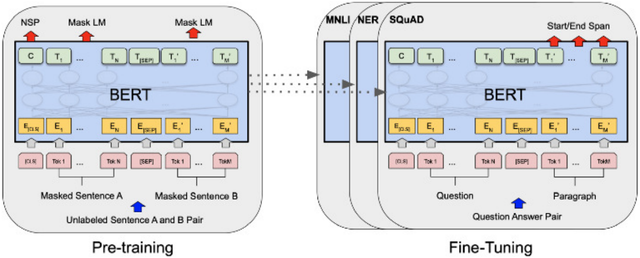
Bidirectional Encoder Representations from Transformers (BERT)



An illustration of Bidirectional Encoder Representations from Transformers (BERT).

- BERT[1] employs a bidirectional Transformer.
- Pre-trained using two unsupervised tasks:
  - Masked Language Modeling (MLM).
  - Next Sentence Prediction (NSP).

[1] Devlin, Chang, Lee, and Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", In *20th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT): Long and Short Papers - Volume 1*, 2019.



An illustration of Bidirectional Encoder Representations from Transformers (BERT).

- BERT[1] employs a bidirectional Transformer.
- Pre-trained using two unsupervised tasks:
  - Masked Language Modeling (MLM).
  - Next Sentence Prediction (NSP).

[1] Devlin, Chang, Lee, and Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", In *20th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT): Long and Short Papers - Volume 1*, 2019.

- Emoji 🙂 $\neq$ emoticons :)
- Emoji = 絵文字, which is a compound word: 絵 (e $\approx$ picture) + 文字 (`moji` $\approx$ written character).
- Invented in Japan in the final years of the 20th century.
- In 2020, approximately one in five tweets included at least one emoji (19.04%).[1]
- In 2017, over 60 million emoji were sent on Facebook and 5 billion emoji were sent on Messenger every day, on average.[2]

[1] Emojipedia, *Emoji Statistics [Updated September 2020]*, 2020, Accessed: 2021-04-26.

[2] Facebook, *5 Billion Emoji Sent Daily on Messenger*, 2017, Accessed: 2021-04-26.

Background

Emoji

- Emoji 🙂 $\neq$ emoticons :)
- Emoji = 絵文字, which is a compound word: 絵 (e $\approx$ picture) + 文字 (`moji` $\approx$ written character).
- Invented in Japan in the final years of the 20th century.
- In 2020, approximately one in five tweets included at least one emoji (19.04%).[1]
- In 2017, over 60 million emoji were sent on Facebook and 5 billion emoji were sent on Messenger every day, on average.[2]

[1] Emojipedia, *Emoji Statistics [Updated September 2020]*, 2020, Accessed: 2021-04-26.

[2] Facebook, *5 Billion Emoji Sent Daily on Messenger*, 2017, Accessed: 2021-04-26.

- Emoji 🙂 $\neq$ emoticons :)
- Emoji = 絵文字, which is a compound word: 絵 (e $\approx$ picture) + 文字 (`moji` $\approx$ written character).
- Invented in Japan in the final years of the 20th century.
- In 2020, approximately one in five tweets included at least one emoji (19.04%).[1]
- In 2017, over 60 million emoji were sent on Facebook and 5 billion emoji were sent on Messenger every day, on average.[2]

[1] Emojipedia, *Emoji Statistics [Updated September 2020]*, 2020, Accessed: 2021-04-26.

[2] Facebook, *5 Billion Emoji Sent Daily on Messenger*, 2017, Accessed: 2021-04-26.

- Emoji 🙂 $\neq$ emoticons :)
- Emoji = 絵文字, which is a compound word: 絵 (e $\approx$ picture) + 文字 (`moji` $\approx$ written character).
- Invented in Japan in the final years of the 20th century.
- In 2020, approximately one in five tweets included at least one emoji (19.04%).[1]
- In 2017, over 60 million emoji were sent on Facebook and 5 billion emoji were sent on Messenger every day, on average.[2]

[1] Emojipedia, *Emoji Statistics [Updated September 2020]*, 2020, Accessed: 2021-04-26.

[2] Facebook, *5 Billion Emoji Sent Daily on Messenger*, 2017, Accessed: 2021-04-26.

Background

Emoji

- Emoji 🙂 $\neq$ emoticons :)
- Emoji = 絵文字, which is a compound word: 絵 (e $\approx$ picture) + 文字 (`moji` $\approx$ written character).
- Invented in Japan in the final years of the 20th century.
- In 2020, approximately one in five tweets included at least one emoji (19.04%).[1]
- In 2017, over 60 million emoji were sent on Facebook and 5 billion emoji were sent on Messenger every day, on average.[2]

😮

[1] Emojipedia, *Emoji Statistics [Updated September 2020]*, 2020, Accessed: 2021-04-26.

[2] Facebook, *5 Billion Emoji Sent Daily on Messenger*, 2017, Accessed: 2021-04-26.

- Dai and Le[1] proposed the supervised fine-tuning step after the unsupervised pre-training. Parameters obtained from the pre-training as a starting point for the supervised training model.
- **Embeddings from Language Models (ELMo)**[2]: contextualized word embedding.
- **Universal Language Model Fine-Tuning (ULMFiT)**[3]: addresses issues of over-fitting and catastrophic forgetting.

[1] Dai and Le, "Semi-Supervised Sequence Learning", In *29th Conference on Neural Information Processing Systems (NIPS)*, 2015.

[2] Peters, Neumann, Iyyer, Gardner, Clark, Lee, and Zettlemoyer, "Deep Contextualized Word Representations", In *16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT): Long Papers - Volume 1*, 2018.

[3] Howard and Ruder, "Universal Language Model Fine-Tuning for Text Classification", In *56th Annual Meeting of the Association for Computational Linguistics (ACL): Long Papers - Volume 1*, 2018.

Background

Related Work

- Dai and Le[1] proposed the supervised fine-tuning step after the unsupervised pre-training. Parameters obtained from the pre-training as a starting point for the supervised training model.
- **Embeddings from Language Models (ELMo)**[2]: contextualized word embedding.
- **Universal Language Model Fine-Tuning (ULMFiT)**[3]: addresses issues of over-fitting and catastrophic forgetting.

---

[1] Dai and Le, "Semi-Supervised Sequence Learning", In *29th Conference on Neural Information Processing Systems (NIPS)*, 2015.

[2] Peters, Neumann, Iyyer, Gardner, Clark, Lee, and Zettlemoyer, "Deep Contextualized Word Representations", In *16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT): Long Papers - Volume 1*, 2018.

[3] Howard and Ruder, "Universal Language Model Fine-Tuning for Text Classification", In *56th Annual Meeting of the Association for Computational Linguistics (ACL): Long Papers - Volume 1*, 2018.

- Dai and Le[1] proposed the supervised fine-tuning step after the unsupervised pre-training. Parameters obtained from the pre-training as a starting point for the supervised training model.
- **Embeddings from Language Models (ELMo)**[2]: contextualized word embedding.
- **Universal Language Model Fine-Tuning (ULMFiT)**[3]: addresses issues of over-fitting and catastrophic forgetting.

[1] Dai and Le, "Semi-Supervised Sequence Learning", In *29th Conference on Neural Information Processing Systems (NIPS)*, 2015.

[2] Peters, Neumann, Iyyer, Gardner, Clark, Lee, and Zettlemoyer, "Deep Contextualized Word Representations", In *16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT): Long Papers - Volume 1*, 2018.

[3] Howard and Ruder, "Universal Language Model Fine-Tuning for Text Classification", In *56th Annual Meeting of the Association for Computational Linguistics (ACL): Long Papers - Volume 1*, 2018.

Background

Related Work

- **Generative Pre-trained Transformer (GPT)**[1]: combines unsupervised pre-training with Transformers, as opposed to Long Short-Term Memory cells.
- **Bidirectional Encoder Representations from Transformers (BERT)**[2]: different training objective (*masked language modeling*).
- **XLNet**[3]: generalized auto-regressive pre-training method.
- **Text-To-Text Transfer Transformer (T5)**[4]: unified text-to-text format where the input and output are always text strings.

[1] Radford, Narasimhan, Salimans, and Sutskever, *Improving Language Understanding by Generative Pre-Training*, 2018.

[2] Devlin, Chang, Lee, and Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", In *20th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT): Long and Short Papers - Volume 1*, 2019.

[3] Yang, Dai, Yang, Carbonell, Salakhutdinov, and Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding", In *33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[4] Raffel, Shazeer, Roberts, Lee, Narang, Matena, Zhou, Li, and Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", *Journal of Machine Learning Research*, 2020.

- **Generative Pre-trained Transformer (GPT)**[1]: combines unsupervised pre-training with Transformers, as opposed to Long Short-Term Memory cells.
- **Bidirectional Encoder Representations from Transformers (BERT)**[2]: different training objective (*masked language modeling*).
- **XLNet**[3]: generalized auto-regressive pre-training method.
- **Text-To-Text Transfer Transformer (T5)**[4]: unified text-to-text format where the input and output are always text strings.

---

[1] Radford, Narasimhan, Salimans, and Sutskever, *Improving Language Understanding by Generative Pre-Training*, 2018.

[2] Devlin, Chang, Lee, and Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", In *20th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT): Long and Short Papers - Volume 1*, 2019.

[3] Yang, Dai, Yang, Carbonell, Salakhutdinov, and Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding", In *33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[4] Raffel, Shazeer, Roberts, Lee, Narang, Matena, Zhou, Li, and Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", *Journal of Machine Learning Research*, 2020.

Background

- **Generative Pre-trained Transformer (GPT)**[1]: combines unsupervised pre-training with Transformers, as opposed to Long Short-Term Memory cells.
- **Bidirectional Encoder Representations from Transformers (BERT)**[2]: different training objective (*masked language modeling*).
- **XLNet**[3]: generalized auto-regressive pre-training method.
- **Text-To-Text Transfer Transformer (T5)**[4]: unified text-to-text format where the input and output are always text strings.

[1] Radford, Narasimhan, Salimans, and Sutskever, *Improving Language Understanding by Generative Pre-Training*, 2018.

[2] Devlin, Chang, Lee, and Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", In *20th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT): Long and Short Papers - Volume 1*, 2019.

[3] Yang, Dai, Yang, Carbonell, Salakhutdinov, and Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding", In *33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[4] Raffel, Shazeer, Roberts, Lee, Narang, Matena, Zhou, Li, and Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", *Journal of Machine Learning Research*, 2020.

- **Generative Pre-trained Transformer (GPT)**[1]: combines unsupervised pre-training with Transformers, as opposed to Long Short-Term Memory cells.
- **Bidirectional Encoder Representations from Transformers (BERT)**[2]: different training objective (*masked language modeling*).
- **XLNet**[3]: generalized auto-regressive pre-training method.
- **Text-To-Text Transfer Transformer (T5)**[4]: unified text-to-text format where the input and output are always text strings.

---

[1] Radford, Narasimhan, Salimans, and Sutskever, *Improving Language Understanding by Generative Pre-Training*, 2018.

[2] Devlin, Chang, Lee, and Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", In *20th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT): Long and Short Papers - Volume 1*, 2019.

[3] Yang, Dai, Yang, Carbonell, Salakhutdinov, and Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding", In *33rd Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[4] Raffel, Shazeer, Roberts, Lee, Narang, Matena, Zhou, Li, and Liu, "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer", *Journal of Machine Learning Research*, 2020.

Agenda

- TweetSentBR[1] was manually annotated.
- 15000 tweets on the TV show domain.
- Three classes: positive, neutral, and negative.
- Predetermined training and test sets.
- About 20% of the samples contain emoji.

[1] Brum and Nunes, "Building a Sentiment Corpus of Tweets in Brazilian Portuguese", In *11th International Conference on Language Resources and Evaluation (LREC)*, 2018.

- TweetSentBR[1] was manually annotated.
- 15000 tweets on the TV show domain.
- Three classes: positive, neutral, and negative.
- Predetermined training and test sets.
- About 20% of the samples contain emoji.

[1] Brum and Nunes, "Building a Sentiment Corpus of Tweets in Brazilian Portuguese", In *11th International Conference on Language Resources and Evaluation (LREC)*, 2018.

- TweetSentBR[1] was manually annotated.
- 15000 tweets on the TV show domain.
- Three classes: positive, neutral, and negative.
- Predetermined training and test sets.
- About 20% of the samples contain emoji.

[1] Brum and Nunes, "Building a Sentiment Corpus of Tweets in Brazilian Portuguese", In *11th International Conference on Language Resources and Evaluation (LREC)*, 2018.

- TweetSentBR[1] was manually annotated.
- 15000 tweets on the TV show domain.
- Three classes: positive, neutral, and negative.
- Predetermined training and test sets.
- About 20% of the samples contain emoji.

[1] Brum and Nunes, "Building a Sentiment Corpus of Tweets in Brazilian Portuguese", In *11th International Conference on Language Resources and Evaluation (LREC)*, 2018.

Datasets

TweetSentBR

- TweetSentBR[1] was manually annotated.
- 15000 tweets on the TV show domain.
- Three classes: positive, neutral, and negative.
- Predetermined training and test sets.
- About 20% of the samples contain emoji.

[1] Brum and Nunes, "Building a Sentiment Corpus of Tweets in Brazilian Portuguese", In *11th International Conference on Language Resources and Evaluation (LREC)*, 2018.

Datasets

TweetSentBR :: Examples

Positive	A fátima fica mais bonita com cabelo curto 💇‍♀️ 😊
Neutral	terminou a entrevista com maluma 🤔
Negative	já 😱 acabouuu nãooo

Average number of words per tweet in TweetSentBR.

Class	Training	Test	Total
Positive	11.37 ± 5.88	11.09 ± 5.66	11.33 ± 5.85
Neutral	11.73 ± 6.12	11.84 ± 6.20	11.74 ± 6.13
Negative	12.91 ± 6.31	13.30 ± 6.32	12.96 ± 6.32
Total	11.92 ± 6.11	11.92 ± 6.07	11.92 ± 6.10

Datasets

TweetSentBR :: Examples

Positive	A fátima fica mais bonita com cabelo curto 💇‍♀️ 😊
Neutral	terminou a entrevista com maluma 🤔
Negative	já 😱 acabouuu nãooo

Average number of words per tweet in TweetSentBR.

Class	**Training**	**Test**	**Total**
Positive	11.37 ± 5.88	11.09 ± 5.66	11.33 ± 5.85
Neutral	11.73 ± 6.12	11.84 ± 6.20	11.74 ± 6.13
Negative	12.91 ± 6.31	13.30 ± 6.32	12.96 ± 6.32
Total	11.92 ± 6.11	11.92 ± 6.07	11.92 ± 6.10

Outline of TweetSentBR.

Class	Training	Test	Total
Positive	5741 (44.2%)	907 (45.1%)	6648
Neutral	3410 (26.3%)	516 (25.7%)	3926
Negative	3839 (29.5%)	587 (29.2%)	4426
Total	12990	2010	15000

Outline of TweetSentBR for tweets containing emoji.

Class	Training	Test	Total
Positive	1688 (64.4%)	274 (66.4%)	1962 (29.5%)
Neutral	379 (14.5%)	65 (15.7%)	444 (11.3%)
Negative	552 (21.1%)	74 (17.9%)	626 (14.1%)
Total	2619 (20.2%)	413 (20.6%)	3032 (20.2%)

Outline of TweetSentBR.

Class	Training	Test	Total
Positive	5741 (44.2%)	907 (45.1%)	6648
Neutral	3410 (26.3%)	516 (25.7%)	3926
Negative	3839 (29.5%)	587 (29.2%)	4426
Total	12990	2010	15000

Outline of TweetSentBR for tweets containing emoji.

Class	Training	Test	Total
Positive	1688 (64.4%)	274 (66.4%)	1962 (29.5%)
Neutral	379 (14.5%)	65 (15.7%)	444 (11.3%)
Negative	552 (21.1%)	74 (17.9%)	626 (14.1%)
Total	2619 (20.2%)	413 (20.6%)	3032 (20.2%)

Top 10 most-frequent emoji of TweetSentBR.

Training		
Unicode	**Emoji**	**Freq.**
U+1F602	😂	1,096
U+1F60D	😍	865
U+02764	❤	737
U+1F44F	👏	518
U+1F62D	😭	282
U+1F622	😢	120
U+1F631	😱	105
U+1F499	💙	93
U+1F3FB	🏻	89
U+02665	♥	75

Test		
Unicode	**Emoji**	**Freq.**
U+1F602	😂	217
U+1F60D	😍	131
U+02764	❤	97
U+1F44F	👏	62
U+1F62D	😭	46
U+1F499	💙	28
U+1F631	😱	25
U+1F622	😢	24
U+1F3B6	🎶	21
U+1F494	💔	13

Datasets

Top 4 most-frequent emoji of TweetSentBR.

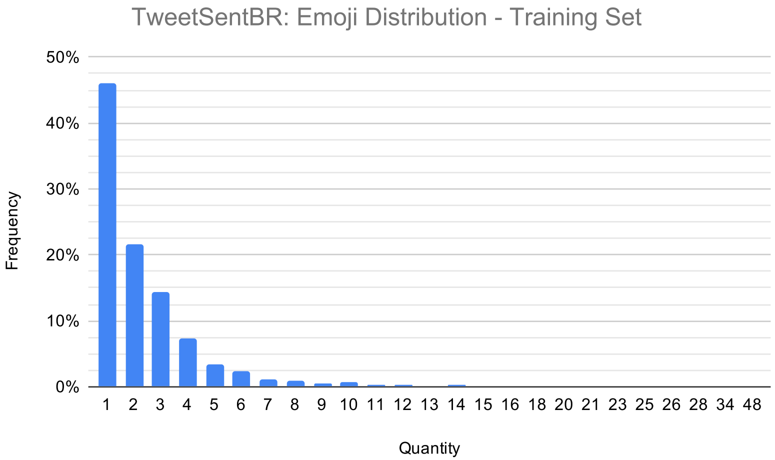
Training		
Unicode	**Emoji**	**Freq.**
U+1F602	😂	1,096
U+1F60D	😍	865
U+02764	❤️	737
U+1F44F	👏	518

Test		
Unicode	**Emoji**	**Freq.**
U+1F602	😂	217
U+1F60D	😍	131
U+02764	❤️	97
U+1F44F	👏	62

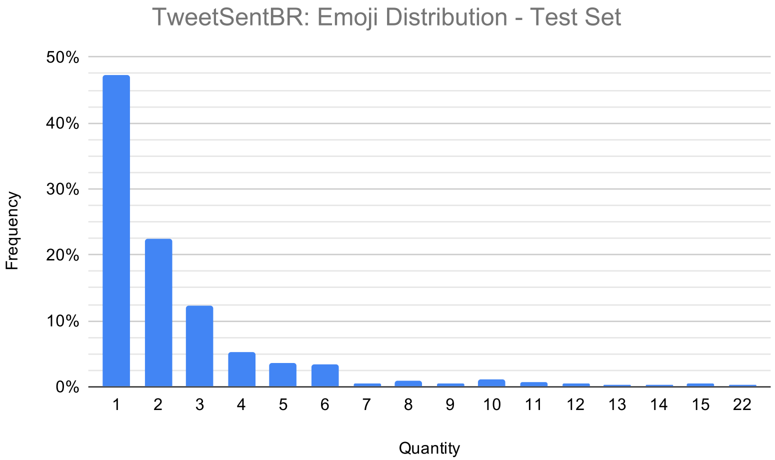
Top 4 most-frequent emoji of Twitter.[1]

Unicode	**Emoji**	**Freq.**
U+1F602	😂	3,103,166,101
U+02764	❤️	1,547,133,201
U+1F60D	😍	1,102,463,335
U+1F62D	😭	1,096,633,626

[1] Emojitracker, *Realtime emoji use on Twitter*, 2013, Accessed: 2020-12-17.



Frequency of emoji in tweets from TweetSentBR – training set.



Frequency of emoji in tweets from TweetSentBR – test set.

- 2000-tweets-BR[1] was also manually annotated.
- 2000 multi-domain tweets.
- Originally four classes: positive, neutral, negative, and mixed.
- Excluding the *mixed* class, we have 1939 tweets.
- 15% of the samples, randomly selected, as test set.
- About 15% of the samples contain emoji.

[1] Vitório, Souza, Teles, and Oliveira, "Investigating Opinion Mining through Language Varieties: a Case Study of Brazilian and European Portuguese tweets", In *11th Brazilian Symposium in Information and Human Language Technology (STIL)*, 2017.

- 2000-tweets-BR[1] was also manually annotated.
- 2000 multi-domain tweets.
- Originally four classes: positive, neutral, negative, and mixed.
- Excluding the *mixed* class, we have 1939 tweets.
- 15% of the samples, randomly selected, as test set.
- About 15% of the samples contain emoji.

[1] Vitório, Souza, Teles, and Oliveira, "Investigating Opinion Mining through Language Varieties: a Case Study of Brazilian and European Portuguese tweets", In *11th Brazilian Symposium in Information and Human Language Technology (STIL)*, 2017.

Datasets

2000-tweets-BR

- 2000-tweets-BR[1] was also manually annotated.
- 2000 multi-domain tweets.
- Originally four classes: positive, neutral, negative, and mixed.
- Excluding the *mixed* class, we have 1939 tweets.
- 15% of the samples, randomly selected, as test set.
- About 15% of the samples contain emoji.

[1] Vitório, Souza, Teles, and Oliveira, "Investigating Opinion Mining through Language Varieties: a Case Study of Brazilian and European Portuguese tweets", In *11th Brazilian Symposium in Information and Human Language Technology (STIL)*, 2017.

- 2000-tweets-BR[1] was also manually annotated.
- 2000 multi-domain tweets.
- Originally four classes: positive, neutral, negative, and mixed.
- Excluding the *mixed* class, we have 1939 tweets.
- 15% of the samples, randomly selected, as test set.
- About 15% of the samples contain emoji.

[1] Vitório, Souza, Teles, and Oliveira, "Investigating Opinion Mining through Language Varieties: a Case Study of Brazilian and European Portuguese tweets", In *11th Brazilian Symposium in Information and Human Language Technology (STIL)*, 2017.

Datasets

2000-tweets-BR

- 2000-tweets-BR[1] was also manually annotated.
- 2000 multi-domain tweets.
- Originally four classes: positive, neutral, negative, and mixed.
- Excluding the *mixed* class, we have 1939 tweets.
- 15% of the samples, randomly selected, as test set.
- About 15% of the samples contain emoji.

[1] Vitório, Souza, Teles, and Oliveira, "Investigating Opinion Mining through Language Varieties: a Case Study of Brazilian and European Portuguese tweets", In *11th Brazilian Symposium in Information and Human Language Technology (STIL)*, 2017.

- 2000-tweets-BR[1] was also manually annotated.
- 2000 multi-domain tweets.
- Originally four classes: positive, neutral, negative, and mixed.
- Excluding the *mixed* class, we have 1939 tweets.
- 15% of the samples, randomly selected, as test set.
- About 15% of the samples contain emoji.

[1] Vitório, Souza, Teles, and Oliveira, "Investigating Opinion Mining through Language Varieties: a Case Study of Brazilian and European Portuguese tweets", In *11th Brazilian Symposium in Information and Human Language Technology (STIL)*, 2017.

Datasets

2000-tweets-BR :: Examples

Positive	O ultimate é lindon 💗 #BTS
Neutral	Quem vive de orgulho morre de saudadeee!! 😜💥
Negative	Não acredito 😨💔💔💔

Average number of words per tweet in 2000-tweets-BR.

Class	**Training**	**Test**	**Total**
Positive	12.50 ± 6.82	10.97 ± 5.88	12.26 ± 6.70
Neutral	11.98 ± 6.79	12.21 ± 6.89	12.01 ± 6.80
Negative	12.84 ± 7.44	12.60 ± 6.85	12.80 ± 7.34
Total	12.30 ± 6.98	12.06 ± 6.68	12.27 ± 6.93

Datasets

2000-tweets-BR :: Examples

Positive	O ultimate é lindon 💗 #BTS
Neutral	Quem vive de orgulho morre de saudadeee!! 😜💥
Negative	Não acredito 😨💔💔💔

Average number of words per tweet in 2000-tweets-BR.

Class	**Training**	**Test**	**Total**
Positive	12.50 ± 6.82	10.97 ± 5.88	12.26 ± 6.70
Neutral	11.98 ± 6.79	12.21 ± 6.89	12.01 ± 6.80
Negative	12.84 ± 7.44	12.60 ± 6.85	12.80 ± 7.34
Total	12.30 ± 6.98	12.06 ± 6.68	12.27 ± 6.93

Outline of 2000-tweets-BR.

Class	**Training**	**Test**	**Total**
Positive	329 (20.0%)	61 (20.9%)	390
Neutral	894 (54.2%)	146 (50.2%)	1040
Negative	425 (25.8%)	84 (28.9%)	509
Total	1648	291	1939

Outline of 2000-tweets-BR for tweets containing emoji.

Class	**Training**	**Test**	**Total**
Positive	79 (30.9%)	14 (34.2%)	93 (23.9%)
Neutral	132 (51.5%)	21 (51.2%)	153 (14.7%)
Negative	45 (17.6%)	6 (14.6%)	51 (10.0%)
Total	256 (15.5%)	41 (14.1%)	297 (15.3%)

Outline of 2000-tweets-BR.

Class	Training	Test	Total
Positive	329 (20.0%)	61 (20.9%)	390
Neutral	894 (54.2%)	146 (50.2%)	1040
Negative	425 (25.8%)	84 (28.9%)	509
Total	1648	291	1939

Outline of 2000-tweets-BR for tweets containing emoji.

Class	Training	Test	Total
Positive	79 (30.9%)	14 (34.2%)	93 (23.9%)
Neutral	132 (51.5%)	21 (51.2%)	153 (14.7%)
Negative	45 (17.6%)	6 (14.6%)	51 (10.0%)
Total	256 (15.5%)	41 (14.1%)	297 (15.3%)

Top 10 most-frequent emoji of 2000-tweets-BR.

Training		
Unicode	**Emoji**	**Freq.**
U+1F602	😂	57
U+02764	❤	52
U+1F60D	😍	40
U+1F644	🙄	20
U+1F3FB	🏻	19
U+1F62D	😭	15
U+1F499	💙	13
U+1F3B6	🎶	12
U+1F494	💔	11
U+1F44C	👌	9

Test		
Unicode	**Emoji**	**Freq.**
U+1F602	😂	14
U+1F62D	😭	14
U+02764	❤	7
U+1F494	💔	5
U+1F44A	👊	5
U+1F44C	👌	3
U+1F60D	😍	2
U+1F64F	🙏	2
U+1F497	💗	2
U+1F62A	😪	2

Datasets

Top 4 most-frequent emoji of 2000-tweets-BR.

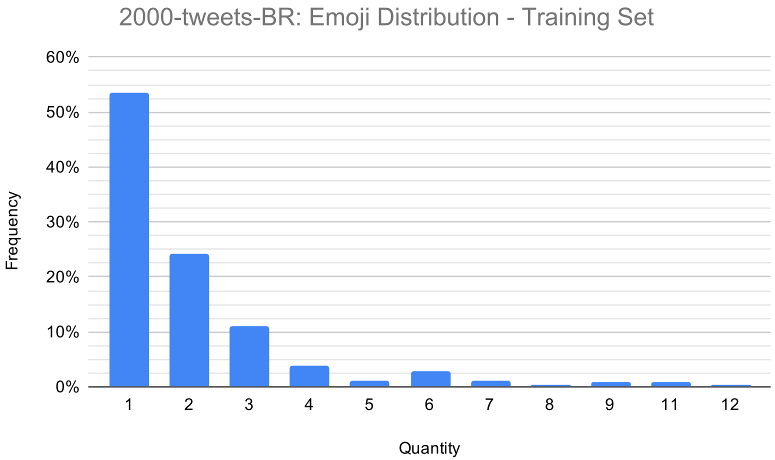
Training		
Unicode	**Emoji**	**Freq.**
U+1F602	😂	57
U+02764	❤️	52
U+1F60D	😍	40
U+1F644	🙄	20

Test		
Unicode	**Emoji**	**Freq.**
U+1F602	😂	14
U+1F62D	😭	14
U+02764	❤️	7
U+1F494	💔	5

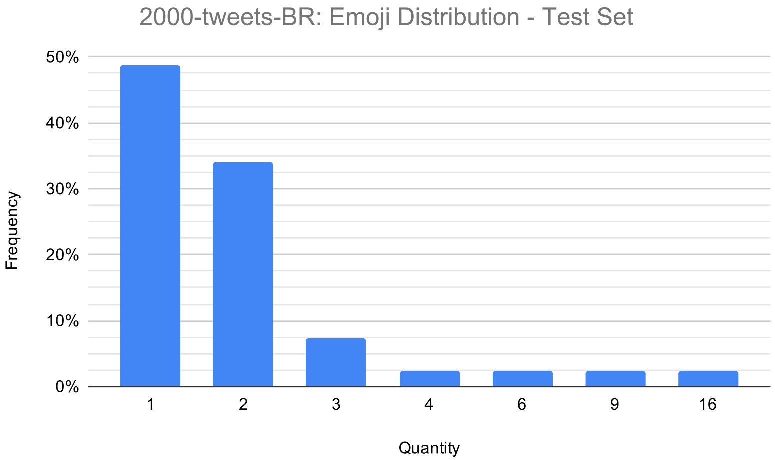
Top 4 most-frequent emoji of Twitter.[1]

Unicode	Emoji	Freq.
U+1F602	😂	3,103,166,101
U+02764	❤️	1,547,133,201
U+1F60D	😍	1,102,463,335
U+1F62D	😭	1,096,633,626

[1] Emojitracker, *Realtime emoji use on Twitter*, 2013, Accessed: 2020-12-17.



Frequency of emoji in tweets from 2000-tweets-BR – training set.



Frequency of emoji in tweets from 2000-tweets-BR – test set.

Agenda

Methodology

Overview

- The core idea behind our proposed methodology is to extract the maximum information possible from emoji to have a richer representation of a piece of text and use that to improve the sentiment classification.

Overview of our method for sentiment classification.

- Use of *BERTimbau*[1] as pre-trained model.
- Corpus of user-generated texts from social media with 89458 samples.
- All of which contain at least one emoji.
- Obtained from social media pages related to TV shows.
- Six different pre-training configurations.

[1] Souza, Nogueira, and Lotufo, "BERTimbau: Pretrained BERT Models for Brazilian Portuguese", In *9th Brazilian Conference on Intelligent Systems (BRACIS)*, 2020.

Methodology

Additional Pre-Training

- Use of *BERTimbau*[1] as pre-trained model.
- Corpus of user-generated texts from social media with 89458 samples.
- All of which contain at least one emoji.
- Obtained from social media pages related to TV shows.
- Six different pre-training configurations.

[1] Souza, Nogueira, and Lotufo, "BERTimbau: Pretrained BERT Models for Brazilian Portuguese", In *9th Brazilian Conference on Intelligent Systems (BRACIS)*, 2020.

Methodology

Additional Pre-Training

- Use of *BERTimbau*[1] as pre-trained model.
- Corpus of user-generated texts from social media with 89458 samples.
- All of which contain at least one emoji.
- Obtained from social media pages related to TV shows.
- Six different pre-training configurations.

[1] Souza, Nogueira, and Lotufo, "BERTimbau: Pretrained BERT Models for Brazilian Portuguese", In *9th Brazilian Conference on Intelligent Systems (BRACIS)*, 2020.

- Use of *BERTimbau*[1] as pre-trained model.
- Corpus of user-generated texts from social media with 89458 samples.
- All of which contain at least one emoji.
- Obtained from social media pages related to TV shows.
- Six different pre-training configurations.

[1] Souza, Nogueira, and Lotufo, "BERTimbau: Pretrained BERT Models for Brazilian Portuguese", In *9th Brazilian Conference on Intelligent Systems (BRACIS)*, 2020.

- Use of *BERTimbau*[1] as pre-trained model.
- Corpus of user-generated texts from social media with 89458 samples.
- All of which contain at least one emoji.
- Obtained from social media pages related to TV shows.
- Six different pre-training configurations.

[1] Souza, Nogueira, and Lotufo, "BERTimbau: Pretrained BERT Models for Brazilian Portuguese", In *9th Brazilian Conference on Intelligent Systems (BRACIS)*, 2020.

Examples of samples from the pre-training corpus.

Linda a Jessica e tem senso de humor. 😊 😍
Quando foi isso? A mulher não ganhou com um nhoque? 🤔
Caramba, que nível.... circo de horrores 😏

- **Masked Language Modeling (MLM)**: the same task used during pre-training of BERT. Random tokens are masked with a probability of 15% and the model is trained to predict those masked tokens.

Example of **MLM**.

Text	**Labels**
Alguém pede pra Jojo ⟨MASK⟩ esse vestido. 🤭 🤣 🤣 ⟨MASK⟩ 🤣 🏃 🏃	trocar 🤣

Methodology

Additional Pre-Training

- **Masked Language Modeling 50% (MLM50)**: similar to the *Masked Language Modeling* configuration, but using a probability of 50% to mask a token.

Example of **MLM50**.

Text	**Labels**
⟨MASK⟩ pede pra ⟨MASK⟩ trocar ⟨MASK⟩ vestido. ⟨MASK⟩ 🤣 ⟨MASK⟩ 🤣 🤣 🏃 ⟨MASK⟩	Alguém Jojo esse 🤭 🤣 🏃

Methodology

Additional Pre-Training

- **All Emoji (All)**: all emoji (and only emoji) are masked and the model is trained to predict those masked emoji.

Example of **All**.

Text	**Labels**
Alguém pede pra Jojo trocar esse vestido. ⟨MASK⟩ ⟨MASK⟩ ⟨MASK⟩ ⟨MASK⟩ ⟨MASK⟩ ⟨MASK⟩ ⟨MASK⟩	🤭 🤣 🤣 🤣 🤣 🏃‍♂️ 🏃‍♂️

- **First Emoji (First)**: the first occurring emoji of a text is masked and the model is trained to predict this masked emoji.

Example of **First**.

Text	**Label**
Alguém pede pra Jojo trocar esse vestido. ⟨MASK⟩ 🤣 🤣 🤣 🤣 🏃‍♂️ 🏃‍♂️	🤭

- **Emoji Masked Language Modeling (EMLM)**: similar to the *Masked Language Modeling* configuration, but only emoji tokens are randomly masked, with a probability of 15%.

Example of **EMLM**.

Text	**Label**
Alguém pede pra Jojo trocar esse vestido. 🤭 🤣 🤣 🤣 🤣 ⟨MASK⟩ 🏃‍♂️	🏃‍♂️

- **Emoji Masked Language Modeling 50% (EMLM50)**: similar to the *Emoji Masked Language Modeling* configuration, but using a probability of 50% to mask a token.

Example of **EMLM50**.

Text	**Labels**
Alguém pede pra Jojo trocar esse vestido. ⟨MASK⟩ 🤣 ⟨MASK⟩ 🤣 🤣 🏃‍♂️ ⟨MASK⟩	🤭 🤣 🏃‍♂️

- Based on Easy Data Augmentation (EDA).[1]
- Four operations:
  - Synonym Replacement.
  - Random Insertion.
  - Random Swap.
  - Random Deletion.
- Two parameters:
  - α: percentage of words in a sample to be changed.
  - n_{aug}: number of augmented samples per original sample.

[1] Wei and Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks", In *24th Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

- Based on Easy Data Augmentation (EDA).[1]
- Four operations:
  - Synonym Replacement.
  - Random Insertion.
  - Random Swap.
  - Random Deletion.
- Two parameters:
  - α: percentage of words in a sample to be changed.
  - n_{aug}: number of augmented samples per original sample.

[1] Wei and Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks", In *24th Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Methodology

Data Augmentation

- Based on Easy Data Augmentation (EDA).[1]
- Four operations:
  - Synonym Replacement.
  - Random Insertion.
  - Random Swap.
  - Random Deletion.
- Two parameters:
  - α: percentage of words in a sample to be changed.
  - n_{aug}: number of augmented samples per original sample.

[1] Wei and Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks", In *24th Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

- Based on Easy Data Augmentation (EDA).[1]
- Four operations:
  - Synonym Replacement.
  - Random Insertion.
  - Random Swap.
  - Random Deletion.
- Two parameters:
  - α: percentage of words in a sample to be changed.
  - n_{aug}: number of augmented samples per original sample.

[1] Wei and Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks", In *24th Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

- Based on Easy Data Augmentation (EDA).[1]
- Four operations:
  - Synonym Replacement.
  - Random Insertion.
  - Random Swap.
  - Random Deletion.
- Two parameters:
  - α: percentage of words in a sample to be changed.
  - n_{aug}: number of augmented samples per original sample.

[1] Wei and Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks", In *24th Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

- Based on Easy Data Augmentation (EDA).[1]
- Four operations:
  - Synonym Replacement.
  - Random Insertion.
  - Random Swap.
  - Random Deletion.
- Two parameters:
  - α: percentage of words in a sample to be changed.
  - n_{aug}: number of augmented samples per original sample.

[1] Wei and Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks", In *24th Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

- Based on Easy Data Augmentation (EDA).[1]
- Four operations:
  - Synonym Replacement.
  - Random Insertion.
  - Random Swap.
  - Random Deletion.
- Two parameters:
  - α: percentage of words in a sample to be changed.
  - n_{aug}: number of augmented samples per original sample.

[1] Wei and Zou, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks", In *24th Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Methodology

Emoji Extraction

- For every tweet, we extract emoji from the text.
- As in the following example:
  - Before:
  - After: que coisa linda amor
- Emoticons[1] are treated as emoji.

[1] `:( =( ;( :-( ;-( :) =) ;) :-) ;-) :D ;D <3 S2`

Methodology

Emoji Extraction

- For every tweet, we extract emoji from the text.
- As in the following example:
  - Before: que coisa linda 😱 😍 awww 😍
  - After: que coisa linda awww
- Emoticons[1] are treated as emoji.

[1] :(=(;(:-(;-(:) =) ;) :-) ;-) :D ;D <3 S2

- For every tweet, we extract emoji from the text.
- As in the following example:
  - Before: | que coisa linda 😱 😍 awww 😍 |
  - After: | que coisa linda awww | | 😱 😍 😍 |
- Emoticons[1] are treated as emoji.

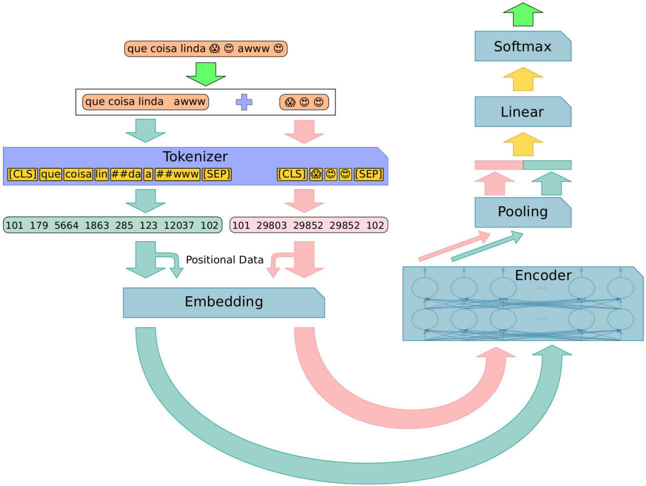
[1] :(=(;(:-(;-(:) =) ;) :-) ;-) :D ;D <3 S2

Methodology

Emoji Extraction

- For every tweet, we extract emoji from the text.
- As in the following example:
  - Before: `que coisa linda 😱 😍 awww 😍`
  - After: `que coisa linda awww` `😱 😍 😍`
- Emoticons[1] are treated as emoji.

[1] `:( =( ;( :-( ;-( :) =) ;) :-) ;-) :D ;D <3 S2`



Our method for sentiment classification.

- Main parameter values:
  - Size of hidden layers $H = 768$.
  - Number of layers $L = 12$.
  - Number of self-attention heads $A = 12$.
- Use of dropout to reduce the overfitting and obtain a better model:
  - Probabilities ranging from 0 to 0.5 in steps of 0.1.

- Main parameter values:
  - Size of hidden layers $H = 768$.
  - Number of layers $L = 12$.
  - Number of self-attention heads $A = 12$.
- Use of dropout to reduce the overfitting and obtain a better model:
  - Probabilities ranging from 0 to 0.5 in steps of 0.1.

Methodology

Model Architecture

- Main parameter values:
  - Size of hidden layers $H = 768$.
  - Number of layers $L = 12$.
  - Number of self-attention heads $A = 12$.
- Use of dropout to reduce the overfitting and obtain a better model:
  - Probabilities ranging from 0 to 0.5 in steps of 0.05

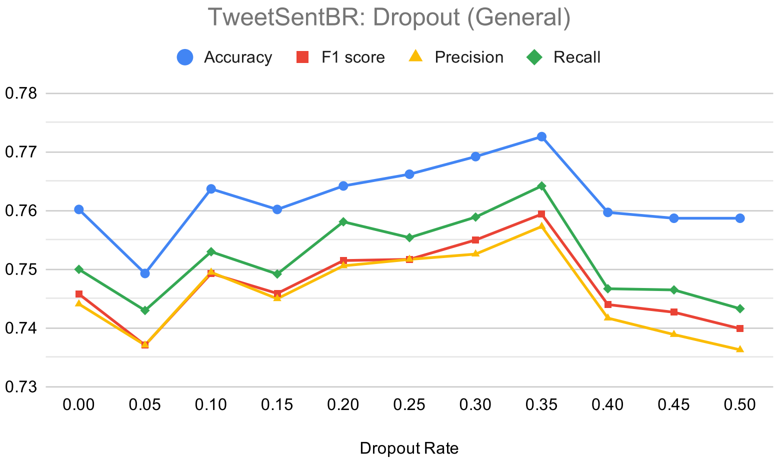
- Main parameter values:
  - Size of hidden layers $H = 768$.
  - Number of layers $L = 12$.
  - Number of self-attention heads $A = 12$.
- Use of dropout to reduce the overfitting and obtain a better model:
  - Probabilities ranging from 0 to 0.5 in steps of 0.05.

Methodology

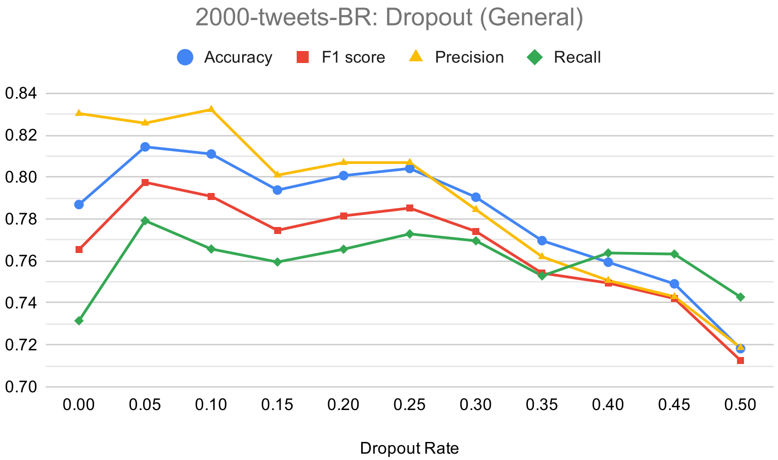
Model Architecture

- Main parameter values:
  - Size of hidden layers $H = 768$.
  - Number of layers $L = 12$.
  - Number of self-attention heads $A = 12$.
- Use of dropout to reduce the overfitting and obtain a better model:
  - Probabilities ranging from 0 to 0.5 in steps of 0.05.

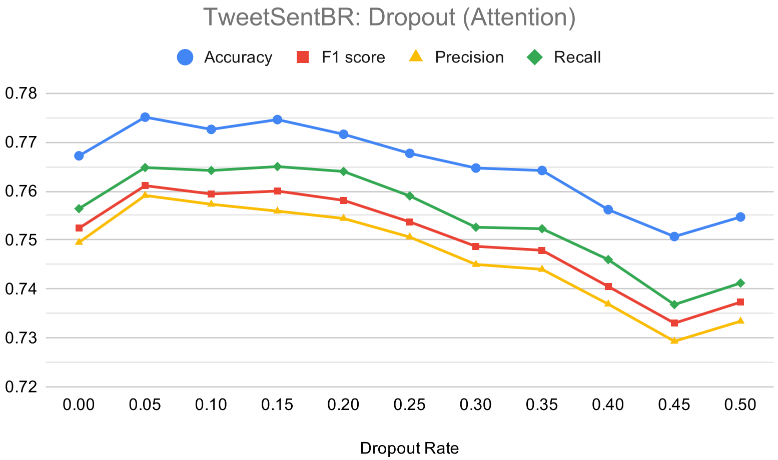
- Main parameter values:
  - Size of hidden layers $H = 768$.
  - Number of layers $L = 12$.
  - Number of self-attention heads $A = 12$.
- Use of dropout to reduce the overfitting and obtain a better model:
  - Probabilities ranging from 0 to 0.5 in steps of 0.05.



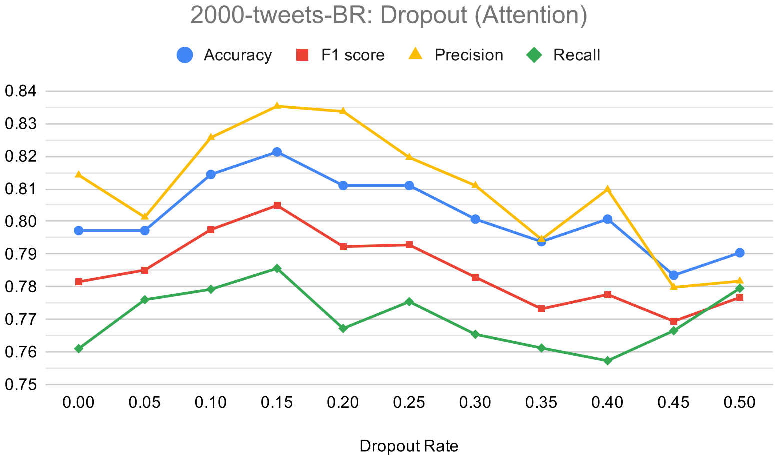
Dropout results for TweetSentBR.



Dropout results for 2000-tweets-BR.



Dropout results for TweetSentBR – self-attention computation.



Dropout results for 2000-tweets-BR – self-attention computation.

- In summary, the best dropout settings for the TweetSentBR dataset are general dropout rate of 35% and self-attention dropout rate of 5%.
- The best settings for the 2000-tweets-BR dataset are general dropout rate of 5% and self-attention dropout rate of 15%.

- In summary, the best dropout settings for the TweetSentBR dataset are general dropout rate of 35% and self-attention dropout rate of 5%.
- The best settings for the 2000-tweets-BR dataset are general dropout rate of 5% and self-attention dropout rate of 15%.

- Maximum of 128 input tokens per sample.
- Stratified 5-fold cross-validation.
- Batch size of 32.
- Maximum number of epochs:
  - 20 for TweetSentBR.
  - 100 for 2000-tweets-BR.
- AdamW optimizer.[1]
  - Initial learning rate of 1×10^{-5}.
  - Weight decay of 0.01.
  - $\beta_1 = 0.9$.
  - $\beta_2 = 0.999$.

[1] Loshchilov and Hutter, "Decoupled Weight Decay Regularization", In *7th International Conference on Learning Representations (ICLR)*, 2019.

Methodology

Training Protocol

- Maximum of 128 input tokens per sample.
- Stratified 5-fold cross-validation.
- Batch size of 32.
- Maximum number of epochs:
  - 20 for TweetSentBR.
  - 100 for 2000-tweets-BR.
- AdamW optimizer.[1]
  - Initial learning rate of 1×10^{-4}.
  - Weight decay of 0.01.
  - $\beta_1 = 0.9$.
  - $\beta_2 = 0.999$.

[1] Loshchilov and Hutter, "Decoupled Weight Decay Regularization", In *7th International Conference on Learning Representations (ICLR)*, 2019.

- Maximum of 128 input tokens per sample.
- Stratified 5-fold cross-validation.
- Batch size of 32.
- Maximum number of epochs:
  - 20 for TweetSentBR.
  - 100 for 2000-tweets-BR.
- AdamW optimizer.[1]
  - Initial learning rate of [illegible]
  - Weight decay of 0.0[illegible]
  - $\beta_1 = 0.9$.
  - $\beta_2 = 0.999$.

[1] Loshchilov and Hutter, "Decoupled Weight Decay Regularization", In *7th International Conference on Learning Representations (ICLR)*, 2019.

Methodology

Training Protocol

- Maximum of 128 input tokens per sample.
- Stratified 5-fold cross-validation.
- Batch size of 32.
- Maximum number of epochs:
  - 20 for TweetSentBR.
  - 100 for 2000-tweets-BR.
- AdamW optimizer.[1]
  - Initial learning rate of 1×10^{-5}.
  - Weight decay of 0.01.
  - $\beta_1 = 0.9$.
  - $\beta_2 = 0.999$.

[1] Loshchilov and Hutter, "Decoupled Weight Decay Regularization", In *7th International Conference on Learning Representations (ICLR)*, 2019.

- Maximum of 128 input tokens per sample.
- Stratified 5-fold cross-validation.
- Batch size of 32.
- Maximum number of epochs:
  - 20 for TweetSentBR.
  - 100 for 2000-tweets-BR.
- AdamW optimizer.[1]
  - Initial learning rate of 1×10^{-5}.
  - Weight decay of 0.01.
  - $\beta_1 = 0.9$.
  - $\beta_2 = 0.999$.

[1] Loshchilov and Hutter, "Decoupled Weight Decay Regularization", In *7th International Conference on Learning Representations (ICLR)*, 2019.

Software:

- Python.
- PyTorch.
- 🤗 Transformers.
- BERTimbau.
- NumPy.
- Pandas.
- NLTK.
- scikit-learn.
- Odysci Media Analyzer.

Hardware:

- Google Colaboratory.
- Processor: 2.2 GHz Intel Xeon.
- Main Memory: 12 GB of RAM.
- Graphics Card: NVidia Tesla P100.
  - Default memory HBM2 of 16 GB.
  - 3584 CUDA cores.

Agenda

- Accuracy:

$$Accuracy(y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{N}(\mathbb{1}(\hat{y}_i = y_i))$$

  - N: number of samples.
  - $\hat{y}_i$: predicted label of the i-th sample.
  - y_i: corresponding true label.
- Balanced accuracy.
- Precision.
- Recall.
- F_1 score.
- Balanced F_1 score.

Experimental Results

- Accuracy.
- Balanced Accuracy:

$$Balanced\ Accuracy = \frac{1}{|C|} \sum_{c \in C} \left(\frac{r_c}{n_c} \right)$$

  - C: set of classes.
  - n_c: number of samples from class c.
  - r_c: number of samples from class c that were predicted correctly.
- Precision.
- Recall.
- F_1 score.
- Balanced F_1 score.

Experimental Results

- Accuracy.
- Balanced Accuracy.
- Precision:

$$Precision(c) = \frac{TP_c}{TP_c + FP_c}$$

  - TP_c: number of “true positives” for class c.
  - FP_c: number of “false positives” for class c.
- Recall.
- F_1 score.
- Balanced F_1 score.

- Accuracy.
- Balanced Accuracy.
- Precision.
- Recall:

$$Recall(c) = \frac{TP_c}{TP_c + FN_c}$$

  - TP_c: number of "true positives" for class c.
  - FN_c: number of "false negatives" for class c.
- F_1 score.
- Balanced F_1 score.

Experimental Results

- Accuracy.
- Balanced Accuracy.
- Precision.
- Recall.
- F_1 score:

$$F_1\ score(c) = 2 \times \frac{Precision(c) \times Recall(c)}{Precision(c) + Recall(c)}$$

$$Macro\text{-}F_1\ score = \frac{1}{|C|} \sum_{c \in C} (F_1\ score(c))$$

  - C: set of classes.
- Balanced F_1 score.

- Accuracy.
- Balanced Accuracy.
- Precision.
- Recall.
- F_1 score.
- Balanced F_1 score:

$$\mathit{Balanced}\ F_1\ \mathit{score} = \frac{1}{\sum_{c \in C} |S_c|} \sum_{c \in C} (|S_c| \times F_1\ \mathit{score}(c))$$

  - C: set of classes.
  - S: set of input samples.
  - S_c: subset of S for the class c.

Pre-training results for TweetSentBR.

Config.	Accuracy	Bal. Acc.	F_1 score	Bal. F_1	Precision	Recall
None	0.7592	0.7476	0.7441	0.7621	0.7425	0.7476
MLM	0.7647	0.7531	0.7495	0.7670	0.7466	0.7531
MLM50	**0.7706**	**0.7576**	**0.7552**	**0.7727**	**0.7532**	**0.7576**
All	0.7567	0.7383	0.7389	0.7589	0.7432	0.7383
First	0.7627	0.7528	0.7487	0.7668	0.7489	0.7528
EMLM	0.7582	0.7445	0.7423	0.7607	0.7431	0.7445
EMLM50	0.7637	0.7500	0.7484	0.7659	0.7482	0.7500

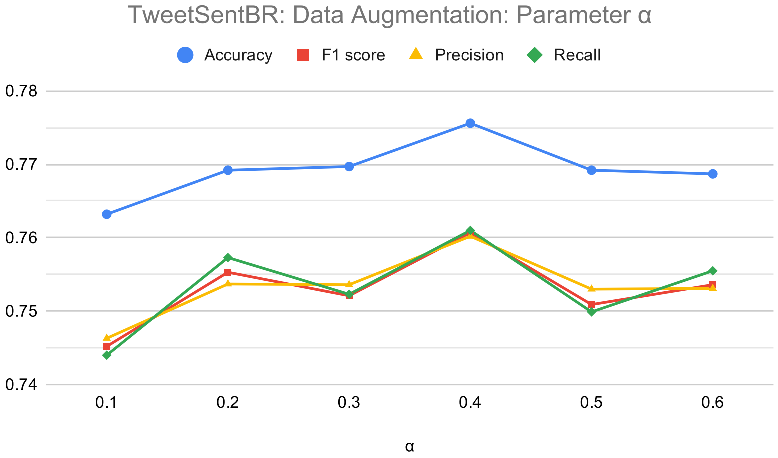
Pre-training results for 2000-tweets-BR.

Config.	Accuracy	Bal. Acc.	F_1 score	Bal. F_1	Precision	Recall
None	**0.8144**	0.7665	**0.7939**	**0.8102**	**0.8416**	0.7665
MLM	0.7972	0.7680	0.7796	0.7952	0.7992	0.7680
MLM50	0.7938	0.7627	0.7758	0.7911	0.7992	0.7627
All	0.7938	0.7414	0.7729	0.7892	0.8310	0.7414
First	0.8041	**0.7871**	0.7896	0.8033	0.7981	**0.7871**
EMLM	0.7938	0.7477	0.7765	0.7900	0.8274	0.7477
EMLM50	0.8041	0.7641	0.7881	0.8013	0.8270	0.7641

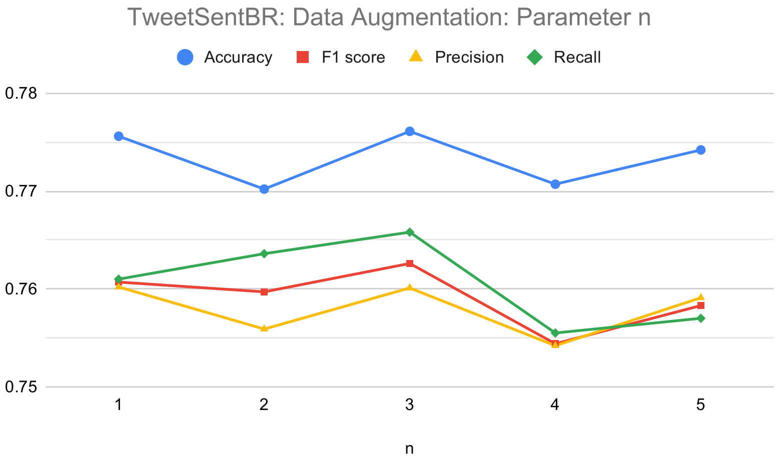
- Finding the best values for the parameters α and n_{aug}.
- α: percentage of words in a sample to be changed.
- n_{aug}: number of augmented samples per original sample.

- Finding the best values for the parameters α and n_{aug}.
- α: percentage of words in a sample to be changed.
- n_{aug}: number of augmented samples per original sample.

- Finding the best values for the parameters α and n_{aug}.
- α: percentage of words in a sample to be changed.
- n_{aug}: number of augmented samples per original sample.



Data augmentation results for TweetSentBR – parameter α.



Data augmentation results for TweetSentBR – parameter n_{aug}.

- In summary, the best data augmentation schema for the TweetSentBR dataset is 3 augmented samples per original sample, with 40% of the words changed.

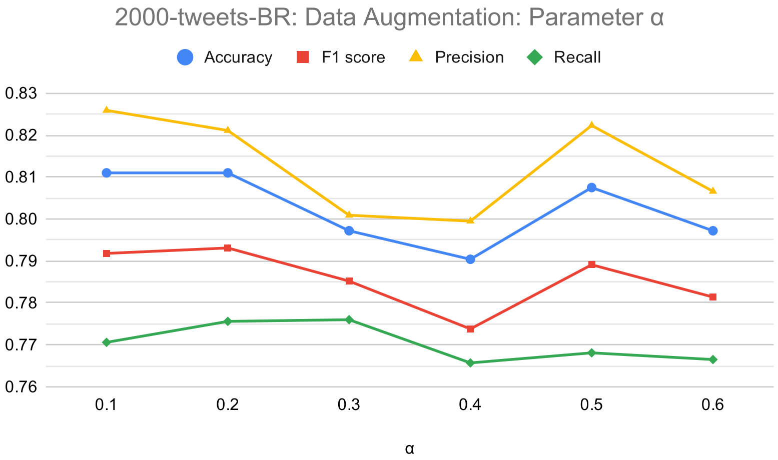
Data augmentation results for TweetSentBR.

Aug.	Accuracy	Bal. Acc.	F_1 score	Bal. F_1	Precision	Recall
No	0.7751	0.7648	0.7611	0.7776	0.7591	0.7648
Yes	**0.7762**	**0.7657**	**0.7625**	**0.7792**	**0.7602**	**0.7657**

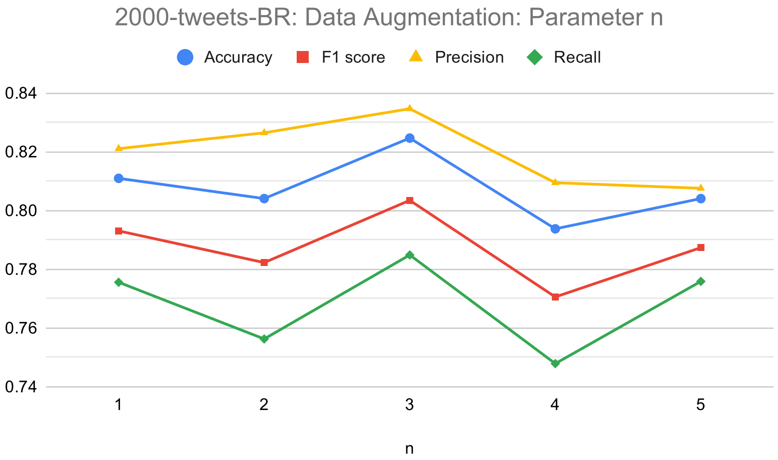
- In summary, the best data augmentation schema for the TweetSentBR dataset is 3 augmented samples per original sample, with 40% of the words changed.

Data augmentation results for TweetSentBR.

Aug.	**Accuracy**	**Bal. Acc.**	**F_1 score**	**Bal. F_1**	**Precision**	**Recall**
No	0.7751	0.7648	0.7611	0.7776	0.7591	0.7648
Yes	**0.7762**	**0.7657**	**0.7625**	**0.7792**	**0.7602**	**0.7657**



Data augmentation results for 2000-tweets-BR – parameter α.



Data augmentation results for 2000-tweets-BR – parameter n_{aug}.

- The best data augmentation schema for the 2000-tweets-BR dataset is 3 augmented samples per original sample, with 20% of the words changed.

Data augmentation results for 2000-tweets-BR.

Aug.	Accuracy	Bal. Acc.	F_1 score	Bal. F_1	Precision	Recall
No	0.8213	**0.7856**	**0.8049**	0.8186	**0.8353**	**0.7856**
Yes	**0.8245**	0.7851	0.8037	**0.8212**	0.8346	0.7851

- No statistical difference according to Wilcoxon signed-rank test.[1]

[1] Wilcoxon, "Individual Comparisons by Ranking Methods", *Biometrics Bulletin*, 1945.

- The best data augmentation schema for the 2000-tweets-BR dataset is 3 augmented samples per original sample, with 20% of the words changed.

Data augmentation results for 2000-tweets-BR.

Aug.	**Accuracy**	**Bal. Acc.**	**F_1 score**	**Bal. F_1**	**Precision**	**Recall**
No	0.8213	**0.7856**	**0.8049**	0.8186	**0.8353**	**0.7856**
Yes	**0.8245**	0.7851	0.8037	**0.8212**	0.8346	0.7851

- No statistical difference according to Wilcoxon signed-rank test.[1]

[1] Wilcoxon, "Individual Comparisons by Ranking Methods", *Biometrics Bulletin*, 1945.

Experimental Results

- The best data augmentation schema for the 2000-tweets-BR dataset is 3 augmented samples per original sample, with 20% of the words changed.

Data augmentation results for 2000-tweets-BR.

Aug.	Accuracy	Bal. Acc.	F_1 score	Bal. F_1	Precision	Recall
No	0.8213	**0.7856**	**0.8049**	0.8186	**0.8353**	**0.7856**
Yes	**0.8245**	0.7851	0.8037	**0.8212**	0.8346	0.7851

- No statistical difference according to Wilcoxon signed-rank test.[1]

[1] Wilcoxon, "Individual Comparisons by Ranking Methods", *Biometrics Bulletin*, 1945.

Fine-tuning results for TweetSentBR.

Model	Accuracy	Bal. Acc.	F_1 score	Bal. F_1	Precision	Recall
Brum and Nunes[1]	0.6462	–	0.5985	–	–	–
Brum and Nunes[2]	–	–	0.6214	–	–	–
Sakiyama et al.[3]	0.6840	–	0.6560	–	–	–
Nascimento[4]	0.7100	–	0.5000	–	–	–
$\text{BERT}_{\text{BASE}}$	0.7468	0.7297	0.7292	0.7546	0.7287	0.7297
Our model	**0.7761**	**0.7658**	**0.7626**	**0.7790**	**0.7601**	**0.7658**

[1] Brum and Nunes, "Building a Sentiment Corpus of Tweets in Brazilian Portuguese", In *11th International Conference on Language Resources and Evaluation (LREC)*, 2018.

[2] Brum and Nunes, "Semi-supervised Sentiment Annotation of Large Corpora", In *13th International Conference on Computational Processing of the Portuguese Language (PROPOR)*, 2018.

[3] Sakiyama, Silva, and Matsubara, "Twitter Breaking News Detector in the 2018 Brazilian Presidential Election using Word Embeddings and Convolutional Neural Networks", In *37th International Joint Conference on Neural Networks (IJCNN)*, 2019.

[4] Nascimento, "Aplicando Ensemble para Classificação de Textos Curtos em Português do Brasil", 2019.

Fine-tuning results for 2000-tweets-BR.

Model	Accuracy	Bal. Acc.	F_1 score	Bal. F_1	Precision	Recall
Vitório et al.[1]	0.6451	–	–	–	–	–
Nascimento[2]	0.6800	–	0.5700	–	–	–
$\text{BERT}_{\text{BASE}}$	0.8110	0.7807	0.7937	0.8086	0.8135	0.7807
Our model	**0.8247**	**0.7849**	**0.8035**	**0.8211**	**0.8347**	**0.7849**

[1] Vitório, Souza, Teles, and Oliveira, "Investigating Opinion Mining through Language Varieties: a Case Study of Brazilian and European Portuguese tweets", In *11th Brazilian Symposium in Information and Human Language Technology (STIL)*, 2017.

[2] Nascimento, "Aplicando Ensemble para Classificação de Textos Curtos em Português do Brasil", 2019.

Fine-tuning results for TweetSentBR – emoji subset.

Model	Accuracy	Bal. Acc.	F_1 score	Bal. F_1	Precision	Recall
$\text{BERT}_{\text{BASE}}$	0.7724	0.5747	0.5631	0.7170	**0.8342**	0.5747
Our model	**0.8208**	**0.7607**	**0.7425**	**0.8193**	0.7311	**0.7607**

Fine-tuning results for 2000-tweets-BR – emoji subset.

Model	Accuracy	Bal. Acc.	F_1 score	Bal. F_1	Precision	Recall
$\text{BERT}_{\text{BASE}}$	0.7073	**0.6587**	**0.6779**	0.7045	0.7146	**0.6587**
Our model	**0.7317**	0.6270	0.6652	**0.7162**	**0.8323**	0.6270

Fine-tuning results for TweetSentBR – emoji subset.

Model	**Accuracy**	**Bal. Acc.**	**F_1 score**	**Bal. F_1**	**Precision**	**Recall**
$\text{BERT}_{\text{BASE}}$	0.7724	0.5747	0.5631	0.7170	**0.8342**	0.5747
Our model	**0.8208**	**0.7607**	**0.7425**	**0.8193**	0.7311	**0.7607**

Fine-tuning results for 2000-tweets-BR – emoji subset.

Model	**Accuracy**	**Bal. Acc.**	**F_1 score**	**Bal. F_1**	**Precision**	**Recall**
$\text{BERT}_{\text{BASE}}$	0.7073	**0.6587**	**0.6779**	0.7045	0.7146	**0.6587**
Our model	**0.7317**	0.6270	0.6652	**0.7162**	**0.8323**	0.6270

Agenda

1. Do emoji, considered alongside their corresponding texts, improve the sentiment classification accuracy?
   - Yes!
2. Can further unsupervised pre-training on in-domain data improve the sentiment classification performance?
3. Considering that unsupervised language representation learning methods are pre-trained on gigabytes of textual data, does data augmentation improve the sentiment classification performance?

Conclusion and Future Work

Research Questions

1. Do emoji, considered alongside their corresponding texts, improve the sentiment classification accuracy?
   - Yes! 😃
2. Can further unsupervised pre-training on in-domain data improve the sentiment classification performance?
3. Considering that unsupervised language representation learning methods are pre-trained on gigabytes of textual data, does data augmentation improve the sentiment classification performance?

Conclusion and Future Work

Research Questions

1. Do emoji, considered alongside their corresponding texts, improve the sentiment classification accuracy?
   - Yes! 😃
2. Can further unsupervised pre-training on in-domain data improve the sentiment classification performance?
   - Since 2000-tweets-BR is a multi-domain dataset, we can only answer for the TweetSentBR dataset, in which case the classification performance was improved.
3. Considering that unsupervised language representation learning methods are pre-trained on gigabytes of textual data, does data augmentation improve the sentiment classification performance?
   - No statistically significant gains were obtained using a simple data augmentation technique.

Conclusion and Future Work

Research Questions

1. Do emoji, considered alongside their corresponding texts, improve the sentiment classification accuracy?
   - Yes! 😃
2. Can further unsupervised pre-training on in-domain data improve the sentiment classification performance?
   - Since 2000-tweets-BR is a multi-domain dataset, we can only answer for the TweetSentBR dataset, in which case the classification performance was improved. 😃
3. Considering that unsupervised language representation learning methods are pre-trained on gigabytes of textual data, does data augmentation improve the sentiment classification performance?

Conclusion and Future Work

Research Questions

1. Do emoji, considered alongside their corresponding texts, improve the sentiment classification accuracy?
   - Yes! 😃
2. Can further unsupervised pre-training on in-domain data improve the sentiment classification performance?
   - Since 2000-tweets-BR is a multi-domain dataset, we can only answer for the TweetSentBR dataset, in which case the classification performance was improved. 😃
3. Considering that unsupervised language representation learning methods are pre-trained on gigabytes of textual data, does data augmentation improve the sentiment classification performance?
   - No statistically significant gains were obtained using a simple data augmentation technique.

Conclusion and Future Work

Research Questions

1. Do emoji, considered alongside their corresponding texts, improve the sentiment classification accuracy?
    - Yes! 😃
2. Can further unsupervised pre-training on in-domain data improve the sentiment classification performance?
    - Since 2000-tweets-BR is a multi-domain dataset, we can only answer for the TweetSentBR dataset, in which case the classification performance was improved. 😃
3. Considering that unsupervised language representation learning methods are pre-trained on gigabytes of textual data, does data augmentation improve the sentiment classification performance?
    - No statistically significant gains were obtained using a simple data augmentation technique. 😟

- Emoji occurrence in the datasets seem to be representative of the usage in social media.
- Emoji are more frequent in positive contexts.
- Our method of sentiment classification can yield good results and surpass the state-of-the-art results for both datasets.
- Our model, which is based on BERT, outperformed BERT itself in most scenarios.

- Emoji occurrence in the datasets seem to be representative of the usage in social media.
- Emoji are more frequent in positive contexts.
- Our method of sentiment classification can yield good results and surpass the state-of-the-art results for both datasets.
- Our model, which is based on BERT, outperformed BERT itself in most scenarios.

- Emoji occurrence in the datasets seem to be representative of the usage in social media.
- Emoji are more frequent in positive contexts.
- Our method of sentiment classification can yield good results and surpass the state-of-the-art results for both datasets.
- Our model, which is based on BERT, outperformed BERT itself in most scenarios.

- Emoji occurrence in the datasets seem to be representative of the usage in social media.
- Emoji are more frequent in positive contexts.
- Our method of sentiment classification can yield good results and surpass the state-of-the-art results for both datasets.
- Our model, which is based on BERT, outperformed BERT itself in most scenarios.

The following papers have been published as results of this research work:

- T.M. Barros, H. Pedrini, Z. Dias. Leveraging Emoji to Improve Sentiment Classification of Tweets. In 36th ACM/SIGAPP Symposium on Applied Computing (SAC) - Knowledge and Language Processing (KLP) track. Gwangju, Republic of Korea, pages 845–852, 2021. Association for Computing Machinery (ACM).
- T.M. Barros, H. Pedrini, Z. Dias. Data-Augmented Emoji Approach to Sentiment Classification of Tweets. In 25th Iberoamerican Congress on Pattern Recognition (CIARP). Porto, Portugal, pages 1–10, 2021. Springer.

Conclusion and Future Work

Publications

The following papers have been published as results of this research work:

- T.M. Barros, H. Pedrini, Z. Dias. Leveraging Emoji to Improve Sentiment Classification of Tweets. In 36th ACM/SIGAPP Symposium on Applied Computing (SAC) - Knowledge and Language Processing (KLP) track. Gwangju, Republic of Korea, pages 845–852, 2021. Association for Computing Machinery (ACM).
- T.M. Barros, H. Pedrini, Z. Dias. Data-Augmented Emoji Approach to Sentiment Classification of Tweets. In 25th Iberoamerican Congress on Pattern Recognition (CIARP). Porto, Portugal, pages 1–10, 2021. Springer.

😃

- More sophisticated method of data augmentation.
- Evaluate datasets in other languages.
- Employ other Transformer models.
- Explore text preprocessing.

- More sophisticated method of data augmentation.
- Evaluate datasets in other languages.
- Employ other Transformer models.
- Explore text preprocessing.

- More sophisticated method of data augmentation.
- Evaluate datasets in other languages.
- Employ other Transformer models.
- Explore text preprocessing.

- More sophisticated method of data augmentation.
- Evaluate datasets in other languages.
- Employ other Transformer models.
- Explore text preprocessing.

Acknowledgments

LIV

Laboratório de Informática Visual

Acknowledgments

LIV

Laboratório de Informática Visual