From GRAPPA to RoBERTa, a huge step froward in inferring sentiments and opinions from Natural Language in Marketing, Applications to BigData from a Covid19 Tweets Collection.

Michel CALCIU
Université Lille, LUMEN, France
mihai.calciu@univ-lille.fr

Jean-Louis MOULINS
Aix Marseille Université, CRETLOG, France
jean-louis.moulins@univ-amu.fr

Francis SALERNO
Université Lille, LUMEN, France
francis.salerno@univ-lille.fr

Abstract

GRAPPA and RoBERTa are acronyms that may sound funny. The first recalls a well-known Italian drink while the other might remember "Dicke Bertha" the famous WW1 canon who was designed to blow-up 3 meters deep concrete walls. In fact, they cover two ways of inferring sentiments and opinions from Natural Language. ***GRAPPA*** is the name we gave to our GeneRal Approach for Parallel Processing Annotations a method that allowed us in a previous paper (Calciu & al, 2021) to significantly reduce computing time with lexicon-based annotations for sentiment analysis on Tweets. ***RoBERTa*** is just a variant of BERT a Transformer based deep learning technology that blue-up the "walls" of Natural Language Processing (NLP). So, the key point of this research is to demonstrate the huge step forward when inferring sentiments and opinions from lexicon-based annotations to AI Transformer based Deep Learning (DL) approaches like BERT. Besides testing the advantages of DL based contextual word embeddings over context ignoring methods that take text as a "bag of words", we review sentiment Analysis of COVID-19-Related Twitter Data as a specialized field due to the imposed conciseness of tweets and the "disaster" represented by the pandemic. The potential of our 2.6 billion tweets collection for transfer learning is discussed with regard to the numerous contemporary state-of-the-art DL pre-trained models and labeled datasets on the subject that are freely available in specialized repositories on the Web.

Introduction

Sentiment analysis may be defined as an ensemble of methods and techniques that automate extraction and analysis of sentiments in a text. Opinion mining (OM) and sentiment analysis are often seen as interchangeable in literature (Pang & Lee, 2008). Information produced by Language may be objective or subjective. The latter can express sentiments and opinions. Sentiments classification is usually done two-way (positive or negative) or three-way (positive, negative, or neutral) (Zimbra & al., 2018). This is also called sentiment polarity.

Due to the restricted length of tweets, Twitter sentiment analysis, appears as a specialized subfield of sentiment analysis as there is no substantial difference between the document and sentence level (Zimbra

& al., 2018; Giachanou & Crestani, 2016).

Twitter is a microblogging platform where people can express their feelings, emotions, and opinions in short messages, called "tweets", that can contain up to 280 characters (Braig & al.,2023). More than 500 million messages per day are posted on Twitter (Braig & al.,2023). This enables information to travel in real-time worldwide. This abundance of information offered promising opportunities for research. Compared to other social media platforms, its uncomplicated access via the Twitter API made it until recently the go-to platform for researchers (Amara & al.,2021; Antonakaki & al.,2021; Buettner & Buettner,2016). Unfortunately, the Twitter API is no longer free since February 2023 due to management changes after the company's takeover by Elon Musk.

Twitter growth was boosted during the COVID-19 pandemic and it has become the place for people to discuss pandemic-related information and share their opinions.

In this paper, we take advantage of several pre-trained DL models and labeled datasets of Covid19 Tweets that are available for research purposes on specialized repositories on the Web, by recovering their text from our huge collection of 2.6 billion tweets, in order to fine-tune those models and compare their performance as sentiment classifiers to the one of lexicon-based methods.

The rest of the paper is organized as follows. After a literature review that focuses on machine learning-based sentiment analysis techniques and compares the best-performing classification algorithms for COVID-19- related Twitter data, we present the methodology of our study and discuss some results. In the end of the paper, we present our conclusions and suggest some further research directions.

Literature review

A recent and rather thorough literature review by Braig & al. (2023) draws on the topics relevant to managing the COVID-19 pandemic and on how sentiment analysis data can support well-informed decisions by gaining insights into the public opinion. It insists on machine learning (ML) classifiers for tweets that are increasingly popular (Kumar & Jaiswal, 2019; Giachanou & Crestani, 2016). Most of this section is a selection from the above-mentioned review.

There are mainly three techniques for sentiment classification: lexicon-based, machine learning and hybrid approaches (Cambria, 2016). Lexicon-based approaches determine the polarity of a text based on the individual polarity of the words that are present in the text (Cambria, 2016). A shortcoming is their inability to deal with semantic rules and linguistic specificities, such as negation, slang, or sarcasm, which is prevalent in natural language texts (Cambria, 2016).

ML-based approaches require training datasets that contain labeled sentiment classes (Jain & Dandannavar, 2016). On these training datasets, classification models are trained and optimized. ML models quantify natural language text based on their feature representation and predict a text’s polarity based on its feature value (Zimbra & al., 2018). Their performance depends on how well the selected features classify text (Gonçalves & al., 2013).

Typical features are bag-of-words (BOW) methods like term frequency-inverse document frequency (TF-IDF) or n-gram, and word embeddings like Word2Vec or GloVe. Word2Vec and GloVe are pre-trained, unsupervised models that create a vector with a cluster of similar words (Bhadane & al., 2015; Ayyub & al.,2020).

The limited availability of required labeled datasets, can be seen as a major downside of ML methods,

especially for novel subjects. Also, large datasets are usually required to optimize the model parameters. The most frequently used classifiers are: Naïve Bayes (NB) (McCallum & Nigam, 1998), Maximum Entropy (ME) (Jaynes, 1957), Support Vector Machine (SVM) (Boser & al., 1992), Logistic Regression (LR), and Random Forest (RF) (Breiman, 2001).

Deep learning (DL) is a subfield of machine learning that mimics the learning process of human brains (Zhang & al., 2018). It is based on the concept of neural networks: Through experience, the composition and strength of neural connections of the brain can be adapted (Jones, 2014). In the last few years, it has become the "Gold Standard" in ML, achieving cutting-edge performance on a variety of cognitive tasks (Alzubaidi & al., 2021). DL has outperformed popular traditional ML techniques in many fields of application, among which NLP (Alzubaidi & al., 2021). Artificial neural networks with many layers can discover new representations (Bengio & al., 2021).

With every level of abstraction, by applying non-linear transformations, the particular representations are elevated to a more abstract level. This mechanism allows filtering for relevant inputs or features for classification tasks. The level of layers reflects the depth of the network (Braig & al., 2023).

In contrast to models with fewer layers, DL models consist of large amounts of neurons and processing layers. DL models can detect structures from unstructured and unlabeled data by applying a back-propagation algorithm (LeCun & al., 2015). Another advantage of deep-learning models is that they do not rely on manually designed feature extraction because the ideal features can be extracted automatically (LeCun & al., 2015). Hence, they do not require the knowledge of a domain expert. The most common model architectures used in DL are Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long-Short Term Memory (LSTM), and Bidirectional Encoder Representations from Transformers (BERT).

BERT is a model developed by an expert team of Google and presented by Devlin et al. (2018). The basis of the model are the transformers. They help understand context and ambiguity by processing a given word in the context of all other words in the sentence rather than being processed individually. In this context, bidirectional means that BERT can read text input in both directions simultaneously rather than sequentially, unlike other language models. Bidirectional learning makes it possible to train with a larger amount of data than with RNNs and CNNs, which require a sequence of data. Pre-training is done using Masked Language Models (MLMs) and Next Sentence Prediction (NSP). The training corpus that has been used is BooksCorpus with 800M words and the English Wikipedia with 2,500M words. Through the pre-training methods, BERT understands the language as it is spoken by predicting a masked word by its context. Because of unlabeled learning, it continues learning as it operates and uses the pre-training as a base layer. In fine-tuning, BERT can be adapted to a specific field through supervised learning by training it with task-specific inputs and outputs (Devlin & al., 2018).

Methodology and Results

We used transfer learning by fine-tuning some of the numerous contemporary SOTA (state-of-the-art) DL pre-trained models and labeled datasets on Twitter that are freely available in specialized repositories on the Web. As the name suggests, a pre-trained model has already been trained on other data; so that the model can be applied to new text to produce high-quality embeddings. Training a high-quality deep learning language model requires a lot of computational resources. Wolf & al. (2019) point out that ***RoBERTa*** was trained on 160 GB of text, and that training this on a typical cloud computing service would cost around 100K USD.

Pre-trained language models can be found in web repositories like HuggingFace. In the R-package “text” they can be set with the model option in the textEmbed() command for example. In this package the

default model is “bert-base-uncased,” and other models can be specified by using their HuggingFace identifier (https://huggingface.co/ models) such as “***roberta-base***” (Liu et al., 2019), “distilbert-base-uncased” (Sanh et al., 2019), or “gpt2” (Radford et al., 2019). Multilingual language models can also represent several languages; multilingual BERT comprises 104 different languages.
These models are usually trained against data which are, reasonably clean (e.g., news articles, blog posts or Wikipedia). They capture word order and context. In order to adapt to specificities of twitter language such transformer-based language models have been trained on Twitter (Barbieri et al., 2020, 2022; Loureiro et al., 2022). Then, these specialized language models have been further fine-tuned for specific NLP tasks, like sentiments analysis using labeled datasets.
There are several Covid19 twitter datasets, that are also freely available on the above-mentioned repositories. They contain the unique identifiers of tweets and labels indicating polarity (positive, negative or eventually neutral). Available tweets are “dehydrated”, meaning that only their unique identifiers are published in those repositories. “Hydrating” or getting the information contained in the tweets, using the Twitter API, has recently become almost impossible for research due to management changes after the company's takeover by Elon Musk. There is no longer free download and only very limited paid downloads available in developer account arrangements.
Luckily this didn't impact our present research and probably nor future research that we intend to do using Covid19 Twitter data as we collected more than 2.6 billion such tweets by “hydrating” a web repository that covered the main period of this pandemic (from January 2020 to February 2023). The potential of this Covid19 tweets collection for research using Machine Learning and especially DL is significant.
For this first version of our research we use resources from TweetNLP (Camacho-Collados & al., 2022), that regroup highly popular models, with thousands of downloads from the Hugging Face model hub every month (Wolf et al., 2020). All language models rely on ***RoBERTa*** (Liu et al., 2019) and XLM-R (Conneau et al., 2020). These models are efficient on standard hardware and free-tiers of cloud computing services, with reasonable speed even without GPU support (Camacho-Collados & al., 2022).
The sentiment analysis task integrated in TweetNLP consists of predicting the sentiment of a tweet with one of the three following labels: positive, neutral or negative. The base dataset for English is the unified TweetEval version of the Semeval-2017 dataset from the task on Sentiment Analysis in Twitter (Rosenthal et al., 2017).
Using several lexicons (dictionaries) that are available for sentiment analysis (see Calciu & al.,2021; Balech & al 2020) we calculated each text’s overall polarity by adding up the individual polarity scores. If more positive than negative words are included in a tweet, the overall polarity is positive. In case the tweet contains an equal amount of positive and negative words, the overall sentiment is neutral (Jain & Dandannavar, 2016). The annotation process that detects and records positive or negative words in a text from a lexicon can become rather slow when data are abundant. In our previous research (Calciu & al., 2021) we describe this process in detail and in order to accelerate it significantly we introduce what we have called the ***GRAPPA*** method we mentioned above. The main advantage of this approach is that no labeled training data is required, and it can be easily adapted to different languages (Drus & Khalid, 2019). This type of approach is considered an unsupervised learning method (Drus & Khalid, 2019). A shortcoming is the inability to deal with semantic rules and linguistic specificities, such as negation, slang, or sarcasm, which is prevalent in natural language texts (Cambria, 2016).
Our results show that all shortcomings of the lexicon-based approaches are largely overcome by the transformer-based DL methods used. Finally, we show that ***RoBERTa*** based DL models really blow-up the walls of NLP as suggested in the title of this research.

Discussion and further research

While our first results proof the superiority of DL over lexicon-based sentiment analysis methods this

comparison should be extended to emotions and eventually to other text classification methods were lexicon-based methods face ML and especially DL ML models.

Also, at this point of our research, we used DL models that were fine-tuned for general twitter language; but Covid19 is a subject that has boosted tremendously exchanges on Twitter and has become research subject on its own. Our further research should specialize these models for Covid19 Tweets and explore other classification methods by using our huge above-mentioned collection of 2.6 billion Covid19 tweets. TweetNLP for example, besides Sentiment Analysis, can perform other tasks like Topic Classification, Irony Detection, Hate Speech Detection, Offensive Language Detection, Emoji Prediction and Emotion Analysis. Other models can deal with misinformation, fake news recognition etc.

Using specialized Covid19 Twitter language models and fine-tuning them for specific NLP tasks, using labeled datasets needs evaluation in an experimental context. Frequently used indicators are accuracy, precision, recall, and F1-Score (Qazi & al., 2017). The accuracy denotes the relationship between correct predictions and the total number of predictions. The correct predictions are comprised of true positives and true negatives.

As mentioned before labeled but “dehydrated” twitter datasets are and will continue to be freely available in web repositories we mentioned before. Our collection of 2.6 billion Covid19 tweets seems rather exhaustive and covers well the main period of the pandemic. Therefor most of the “dehydrated” tweets from the mentioned datasets can be found in our collection and in this way their text and other important information can be recovered, despite today's restrictions on Twitter data. This collection can support many important future research subjects in marketing and management in fields like health and pharmaceutical marketing (Zaynab, 2024), disaster or crisis management (Balech & al. 2022) and so on.

The collection can also be explored and offer new opportunities using the Computational Grounded Theory (Nelson, 2020), a new research paradigm, which brings more interpretative liberty by articulating the positivist paradigm of the algorithm and the inductive, grounded paradigm of the researcher, following 3 steps: 1) unsupervised models identify themes within a corpus, 2) the researcher interprets the data through in-depth reading, and finally 3) NLP tools validate the classification (Beriche & al, 2024).

References

Alzubaidi L., Zhang J., Humaidi A. J., Al-Dujaili A., Duan Y., Al-Shamma O., Santamaría J., Fadhel M. A., Al-Amidie M. & Farhan L. (2021), "Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, pp. 1–74, Mar.

Amara A., Taieb M. A. H. & Aouicha M. B. (2021), "Multilingual topic mod- eling for tracking COVID-19 trends based on Facebook data analysis," *Appl. Intell*., vol. 51, no. 5, pp. 3052–3073, Feb.

Antonakaki D., Fragopoulou P. & Ioannidis S. (2021), "A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks," *Exp. Syst. Appl*., vol. 164, Feb. 2021, Art. no. 114006.

Ayyub K.,Iqbal S.,Munir E.U.,Nisar M.W. & AbbasiM. (2020), "Exploring diverse features for sentiment quantification using machine learning algorithms," *IEEE Access*, vol. 8, pp. 142819–142831.

Balech S., Benavent C., Calciu M. & Monot. J. (2022) "Le masque, figure polaire de la crise de la Covid-19 : une exploration par NLP du flux des conversations Twitter (février-mai 2020).", Marché et organisations, 1/ 43, p. 151-187

Bengio Y., LeCun Y. & Hinton G.E. (2021),"Deep learning for AI," *Commun. ACM*, vol. 64, pp. 58–65, Jun.

Beriche A., Crié D. & Calciu M (2024), "Une Approche Computationnelle Ancrée : Étude de cas des tweets du challenge #Movember en prévention de santé masculine ", accepté pour Décisions Marketing (numéro spécial Marketing and Artificial Intelligence)

Breiman L. (2001), "Random forests," *Mach. Learn*., vol. 45, no. 1, pp. 5–32, Oct.

Bhadane C., Dalal H. & Doshi H. (2015), "Sentiment analysis: Measuring opinions," *Proc. Comput. Sci*., vol. 45, pp. 808–814, Jan.

Boser B. E., Guyon I. M. & Vapnik V. N. (1992), "A training algorithm for optimal margin classifiers," in *Proc. 5th Annu. Workshop (COLT)*, Jul., pp. 144–152.

Buettner R. & Buettner K. (2016), "A systematic literature review of Twitter research from a socio-political revolution perspective," in *Proc. 49th Hawaii Int. Conf. Syst. Sci. (HICSS)*, Jan. 2016, pp. 2206–2215.

Calciu M., Benavent C., Moulins J-L. & Salerno F. (2021) "The GRAPPA method for accelerating annotations on Big consumer opinion datasets. Applications to sentiment modeling on COVID19 Lockdown Tweets and Amazon", The 20th International Marketing Trends Conference, Venice, January 14-16.

Camacho-Collados J., Rezaee K., Riahi T., Ushio A., Loureiro D., Antypas D., Boisson J., Espinosa-Anke L., Liu F., Martínez-Cámara E., Medina G., Buhrmann T., Neves L. & Barbieri F. (2022), TweetNLP: Cutting-Edge Natural Language Processing for Social Media, Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38 – 49, Dec. 7-11.

Cambria A. (2016), "Affective computing and sentiment analysis," *IEEE Intell. Syst*., vol. 31, no. 2, pp. 102–107, Mar./Apr. 2016.

Devlin J., Chang M.-W., Lee K. & Toutanova K. (2018), "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Oct., pp. 4171–4186.

Drus Z. & Khalid H. (2019), "Sentiment analysis in social media and its application: Systematic literature review," *Proc. Comput. Sci*., vol. 161, pp. 707–714, Jan.

Giachanou A. & Crestani F., (2016), "Like it or not: A survey of Twitter sentiment analysis methods," *ACM Comput. Surv*., vol. 49, no. 2, pp. 1–41, Jun.

Gonçalves P., Araújo M., Benevenuto F. & Cha M. (2013), "Comparing and combining sentiment analysis methods," in Proc. *1st ACM Conf. Online Social Netw*., Oct., pp. 27–38.

Jain A. P. & Dandannavar P. (2016), "Application of machine learning tech- niques to sentiment analysis," in Proc. *2nd Int. Conf. Appl. Theor. Comput. Commun. Technol. (iCATccT)*, Jul., pp. 628–632.

Jaynes E. T. (1957) , "Information theory and statistical mechanics," *Phys. Rev*., vol. 106, no. 4, pp. 620–630.

Jones N. (2014),"Computer science: The learning machines," Nature, vol.505, no. 7482, pp. 146–148, Jan.

Kumar A & Jaiswal A. (2019), "Swarm intelligence based optimal feature selection for enhanced predictive sentiment accuracy on Twitter," *Multimedia Tools Appl*., vol. 78, no. 20, pp. 29529–29553, Oct.

LeCun Y., Bengio Y. & Hinton G. E. (2015), "Deep learning," *Nature*, vol. 521, pp. 436–444, Dec.

McCallum A. & Nigam K. (1998), "A comparison of event models for naive Bayes text classification," in Proc. AAAI Workshop Learn. Text Categorization, vol. 752, no. 1, pp. 41–48.

Pang B. and Lee L. (2008), "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr*., vol. 2, nos. 1–2, pp. 1–135, Jan.

Qazi A., Raj R. G., Hardaker G. & Standing C. (2017), "A systematic literature review on opinion types and sentiment analysis techniques," *Internet Res.,* vol. 27, no. 3, pp. 608–630, Jun.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L. & Polosukhin I. (2017), "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst*., vol. 30, pp. 5998–6008.

Zaynab Z. (2024) " Big pharma's electronic word of mouth during Covid-19", accepted at the *23rd International Marketing Trends Conference*, Venice, January 18-20

Zhang L., Wang S. & Liu B. (2018),"Deep learning for sentiment analysis: A survey," *WIREs Data Mining*

and Knowledge Discovery, vol. 8, p. e1253, Mar.
Zimbra D., Abbasi A., Zeng D. and Chen H. (2018), "The state-of-the-art in Twitter sentiment analysis," *ACM Trans. Manage. Inf. Syst*., vol. 9, no. 2, pp. 1–29, Jun.