INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO

MESTRADO EM ENGENHARIA DE INTELIGÊNCIA ARTIFICIAL

Near Real-Time Sentiment and Topic Analysis of Sport Events

MIGUEL FERRAZ BARBOSA SOARES DE ALBERGARIA

julho de 2022

Near Real-Time Sentiment and Topic Analysis of Sport Events

Miguel Ferraz Barbosa Soares de Albergaria

Aluno nº: 1170551

Dissertação para obtenção do Grau de
Mestre em Engenharia de Inteligência Artificial

Orientador: Doutor Luiz Felipe Rocha de Faria, Professor Coordenador do Instituto Superior de Engenharia do Instituto Politécnico do Porto

Co-orientadora: Doutora Maria Goreti Carvalho Marreiros, Professora Coordenadora com Agregação do Instituto Superior de Engenharia do Instituto Politécnico do Porto

Júri:

Presidente:

Doutor Carlos Fernando da Silva Ramos, Professor Coordenador Principal do Instituto Superior de Engenharia do Instituto Politécnico do Porto

Vogais:

Doutor Alberto Manuel Brandão Simões, Checkmarx

Doutor Luiz Felipe Rocha de Faria, Professor Coordenador do Instituto Superior de Engenharia do Instituto Politécnico do Porto

Porto, junho 2022

Resumo

Os padrões de consumo de media, têm vindo a mudar para um paradigma de ecrãs múltiplos, onde, através de multitasking, os telespetadores podem pesquisar informações adicionais sobre o evento que estão a assistir, bem como partilhar a sua perspetiva do evento. As indústrias do setor audiovisual e multimédia, no entanto, não estão a aproveitar esta oportunidade, falhando em fornecer às equipas desportivas e aos responsáveis pela produção audiovisual uma visão sobre a perspetiva dos consumidores finais dos eventos desportivos.

Como resultado desta oportunidade, este documento foca-se em apresentar o desenvolvimento de uma ferramenta de análise de sentimento e uma ferramenta de análise de tópicos para a análise, em perto de tempo real, de conteúdo das redes sociais relacionado com eventos esportivos e publicado durante a transmissão dos respetivos eventos, permitindo assim, em perto de tempo real, perceber o sentimento dos espectadores e os tópicos mais falados durante cada evento.

Palavras-chave: Processamento de Linguagem Natural, Análise de Sentimentos, Análise de Tópicos

Abstract

Sport events’ media consumption patterns have started transitioning to a multi-screen paradigm, where, through multitasking, viewers are able to search for additional information about the event they are watching live, as well as contribute with their perspective of the event to other viewers. The audiovisual and multimedia industries, however, are failing to capitalize on this by not providing the sports’ teams and those in charge of the audiovisual production with insights on the final consumers perspective of sport events.

As a result of this opportunity, this document focuses on presenting the development of a near real-time sentiment analysis tool and a near real-time topic analysis tool for the analysis of sports events’ related social media content that was published during the transmission of the respective events, thus enabling, in near real-time, the understanding of the sentiment of the viewers and the topics being discussed through each event.

Keywords: Natural Language Processing, Sentiment Analysis, Topic Analysis

Acknowledgement

I would like to start by expressing my deepest gratitude towards Professors Goreti Marreiros and Luiz Faria, which through the roles of supervisors and advisors, guided me in the development of the work hereby presented. I also thank them for giving me the opportunity to join the GECAD research center and take part in the development of the PLAYOFF project.

I thank GECAD for providing me with the necessary conditions to develop this work successfully and for the good working environment.

And I thank all those who helped me, directly or indirectly, throughout my entire academic path, with special emphasis on my parents and closest friends.

Index

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
API	Application Programming Interface
BoW	Bag-of-Words
CBOW	Continuous Bag-of-Words
CNN	Convolutional Neural Network
EDA	Exploratory Data Analysis
ERDF	European Regional Development Fund
IPP	Instituto Politécnico do Porto
ISEP	Instituto Superior de Engenharia do Porto
KDD	Knowledge Discovery in Databases
KDT	Knowledge Discovery in Text
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSTM	Long Short-Term Memory
MAU	Monthly Active User
ML	Machine Learning
MLM	Masked-Language Modelling
NLP	Natural Language Processing
NMF	Non-Negative Matrix Factorization
NN	Neural Network
NSP	Next Sentence Prediction
PLM	Pre-trained Language Model
PMI	Pointwise Mutual Information
RNN	Recurrent Neural Network

SOTA	State of the Art
SVD	Singular Value Decomposition
SVM	Singular Value Decomposition
TF-IDF	Term Frequency-Inverse Document Frequency
URL	Uniform Resource Locator

1 Introduction

This document intends to detail the Natural Language Processing (NLP) component of the "PLAYOFF" project, carried out at the Research Group on Intelligent Engineering and Computing for Advanced Innovation and Development (GECAD[1]), as part of the consortium formed by MOG Technologies[2], leading and business promoter, and Instituto Superior de Engenharia do Porto (ISEP[3]), as a research partner of the Portuguese National Research and Innovation System. With the latter being home to GECAD and belonging to Instituto Politécnico do Porto (IPP[4]).

On this chapter, the scope and objectives of the project are contextualized. The approaches used for the development of the project, along with its key findings and contributions are provided. And, at last, the structure used for this document is presented.

1.1 Context

The proposed project, dubbed "PLAYOFF" after the project title "Personalized LAYered multi-source content - Optimized with data Fusion topologies for sports Fans" and co-financed by Portugal 2020 "European Regional Development Fund" (ERDF) through the 2020 Northern Regional Operational Program, intends to provide a multimodal reactive media transmission environment, aimed at television stations and clubs. However, with the author only being responsible for the development of the NLP side of the project, this document will focus on the

[1] GECAD Website - https://www.gecad.isep.ipp.pt
[2] MOG Technologies Website - https://www.mog-technologies.com
[3] ISEP Website - https://www.isep.ipp.pt
[4] IPP website - https://www.ipp.pt

development of the sentiment and topic analysis modules of a near real-time data fusion and analysis framework for sporting events, which aims to streamline the process of near real-time collection and analysis of social media text content related to sport events.

1.2 Problem Description

In recent years, with the increase in online data availability, in large part due to the rise in social media usage (Poushter et al., 2018; We Are Social; DataReportal; Hootsuite, 2021), media consumption patterns have started transitioning to a multi-screen paradigm, where, through multitasking, viewers are able to search for additional information about the event they are watching live, as well as contribute with their perspective of the event to other viewers. As proven by a Google and Ipsos Connect United States' sports viewers study, 80% of sports fans aged 18 to 54 answered that they use a computer or smartphone while watching live sports on TV to gather additional information about the events and its players, or to message other fans (Google & Ipsos Connect, 2017).

Furthermore, this increase in online data availability, particularly regarding individuals' social media data, has also resulted in significant changes to a diverse spectrum of fields, other than the information and communications technologies' field, such as the one of marketing, where new approaches have been developed to incorporate the information gathered from the aforementioned data into the process of strengthening the relationship between brands and their customers (Saravanakumar & SuganthaLakshmi, 2012), as well as improving the process of finding new customers (Iyer et al., 2005).

It is also imperative to realize that the recent growth in the use of social networks and, consequently, the increase in the amount of publicly available online information, is due, in large part, to the global rise in internet coverage and smartphone ownership (Delaporte & Bahia, 2021). Moreover, smartphones, given their increasingly better capabilities, coupled with the growingly superior worldwide internet speeds (Delaporte & Bahia, 2021), have provided its users with capabilities not long ago only available to industry professionals. Such is the case with the audiovisual content creation industry, which is now accessible to anyone with a smartphone camera and internet access, enabling ordinary people to take part in the process that is the capture of video of events and their ambience.

The audiovisual and multimedia industries, however, are failing to capitalize on these opportunities by not providing an ecosystem able to interconnect all partners of the value of chain of sporting events, namely those in charge of audiovisual production, clubs, and the final consumers.

1.2.1 Objectives

Within the scope of the “PLAYOFF” project and in light of the Problem Description section, it is, therefore, critical to develop a solution capable of gathering and analyzing textual data from social networks, thereby providing television stations and clubs with access to information that may be useful in helping understand the audiences’ perception of sports events and their clubs and, consequently, aiding the further development of their marketing strategies.

As a result, bearing in mind that this solution must be easily integrated into the mentioned project, it is intended the development of a framework which can collect, aggregate, and process text data from a variety of sources, in as close to real-time as possible, while still achieving a high-level performance. Whether these sources be social media, social media content aggregators, or third-party tools, which already gather and process the social networks’ data.

The main objective of the proposed fusion and analysis framework is, therefore, to provide the “PLAYOFF” project with the necessary tools for the near real-time collection and analysis of real-time sports events related text content from social networks. With this document, among the several analysis tools, focusing on the sentiment and topic analysis tools.

With that said, due to time restrictions and as per MOG Technologies' request, the intended framework, while planned to be capable of gathering and aggregating data from several sources, will initially only require one data source, with the addition of more as future work. Moreover, also due to time constraints and considering MOG Technologies’ main target audience, the framework, as a first step, will only need to be able to analyze English text content related to football, with support for additional sports and languages to be added as future development.

Thus, in short, the main tasks necessary to carry out the successful development of the sentiment and topic analysis modules and, consequently, the framework are the following:

- Study of existing social networks, regarding their availability of real-time sport events related text content, APIs, and users' demographics data;
- Study of existing third-party tools for the collection and analysis of social media's text content;
- Study of the literature of the sentiment and topic analysis tasks;
- Experimentation of several pre-processing and processing techniques for the sentiment and topic analysis of the social media's sports related text content;
- Evaluation and comparison of the performance of different models and techniques;

1.2.2 Approach and results

With the ever-growing number of models and techniques being introduced in the in the field of artificial intelligence (AI), many of which improving the state of the art of their respective subfields, it is imperative that a literature review is done prior to the development of any AI tool, in order to help with the identification of the most promising models and techniques to be used for the task at hand.

As such, considering that the quality of the final solution depends on the quality of the literature study done, for both the sentiment and topic analysis, it was decided for the use of only English sources, which contributed significantly to their fields of study, either by having been an important steppingstone in their field or by answering one of the following research questions:

- What machine learning models or techniques can be used for near real-time sentiment analysis?
- How can sentiment analysis models or techniques be optimized?
- What machine learning models or techniques can be used for near real-time topic analysis?
- What topic models can learn based on a set of predefined topics?

From this literature review methodology and considering that the best dataset found had sentiment labels for all its 6.3 million data points, but no topic labels, it was possible for the identification of several promising models and techniques for both proposed tasks. Such as, for the sentiment analysis, the DistilBERT (Sanh et al., 2019), DistilRoBERTa (Hugging Face, 2019) and knowledge embedding of these pre-trained supervised models (Ostendorff et al., 2019), and, for the topic analysis, the JoSH (Meng et al., 2020) and WeSHClass (Meng et al., 2019) weakly-supervised models.

As for the methodology used for the experimentation of techniques, regarding the pre-processing, it is defined a base combination of steps with subsequent experiments, either adding a new pre-processing step to a previous experiment or replacing an already tested step. For the tuning of hyperparameters a base experiment is also defined, however with each subsequent experiment only changing one hyperparameter value at a time. Regarding the experiments for the selection of the final model, each model is simply tested using the same pre-processed data, thus enabling their comparison.

As such, following the experiments methodology on both the sentiment and topic analysis tasks, it was concluded that for the pre-processing, topic analysis benefits more from less complex data than the sentiment analysis. As for the final models it was concluded that for the sentiment and topic analysis, the best models were, respectively, the knowledge embedded DistilRoBERTa and the WeSHClass.

1.2.3 Contributions

The development of the near real-time sentiment and topic analysis tools, although primarily aiming to contribute to the “PLAYOFF” project by providing insights on the final consumers perspective of sport events, also contributes to their respective fields by providing:

- An analysis and comparison of social medias regarding their real-time availability of sports-related text content, APIs and users’ demographics;
- An analysis of existing tools for the sentiment and topic analysis of tweets;
- A literature review of the sentiment and topic analysis fields;
- An exploratory data analysis of a dataset of tweets for the sentiment and topic analysis;
- The results of the pre-processing and processing experiments of the most promising techniques gathered from the literature review;

As such, whether or not the development of the tools proves to be successful, this work contributes to the sentiment and topic analysis fields by providing the readers with a comprehensive guide on what works, what does not work, and what could be improved for better results, thus allowing for this work to be used as a steppingstone for future works.

1.3 Document Structure

This last subsection of the Introduction chapter intends to briefly expose the structure and content of the document, which is divided into five main chapters: Introduction, Social Networks and Text Mining, Sentiment and Topic Analysis, Experimentation of Techniques and Conclusion.

On the first chapter, which is the Introduction, the scope and objectives of the project are contextualized. The approaches used for the development of the project, along with its key findings and contributions are provided. And, at last, the structure used for this document is presented.

Following the Introduction comes the Social Networks and Text Mining chapter, where the results relative to the three first main tasks presented in the Objectives section of the Introduction chapter are described. Thus, this chapter starts by introducing the social networks considered for the near real-time data collection, comparing each other, and justifying the chosen option. Then, the study results of the state of the art of the sentiment and topic analysis tasks are provided, along with the related works and their contributions to the current state of the field or fields in question. To finalize, a comparison between the related works and the data fusion and analysis framework is presented, coupled with a conclusion on the most promising sentiment and topic analysis SOTA techniques presented.

On the third chapter, titled Sentiment and Topic Analysis, the dataset used for the training of the models is presented, along with its exploratory data analysis. Additionally, the procedures taken to maintain data security, coupled with the potential risks or ethical violations of the tools to be developed are also detailed. At last, the most promising pre-processing and processing techniques, for, both, the sentiment and topic analysis are presented.

On the fourth chapter, the evaluation metrics considered for the measurement of the performance of the models are presented, along with the experiments done to compare the different pre-processing and processing techniques for, both, the sentiment and topic analysis. Additionally, the final trained models for both tasks are also presented, along with a first unsuccessful approach for the de the topic analysis task.

On the fifth, and final, chapter a summary of the developed work is presented, along with the conclusions reached from it. Additionally, the limitations of the sentiment and topic analysis tool are presented along with possible future development paths.

2 Social Networks and Text Mining

In this chapter, the results relative to the three first main tasks presented in the Objectives section are described. Thus, this chapter starts off by introducing the social networks considered for the near real-time data collection, comparing each other, and justifying the chosen option. Then, the study results of the state of the art (SOTA) of the sentiment and topic analysis tasks are provided, along with the related works and their contributions to the current state of the field or fields in question. At last, a comparison between the related works and the data fusion and analysis framework is presented, coupled with a conclusion on the most promising sentiment and topic analysis SOTA techniques presented.

2.1 Social Networks

First introduced in (Barnes, 1954) to describe the relationships between pairs of persons in a society, the term "social network" has since evolved beyond its original fields of study, anthropology (Mitchell, 1974) and sociology (Scott, 2002), into the field of software, earning with it, a new meaning, significance and popularity. Nowadays, this term, once unknown and unused by most people, has become a symbol of the digital age, representing a new type of software platforms aimed at allowing their users to interact and present themselves to audiences that value "user-generated content and the perception of interaction with others" (Carr & Hayes, 2015).

Furthermore, these platforms, which have now reached over 4.5 billion active users (Kemp, 2021b), as depicted in Figure 1, have grown to be such a significant aspect of today's world, that for a large number of industries they are already recognized as a tool that can help further advance their strategies (Mayfield, III, 2011).

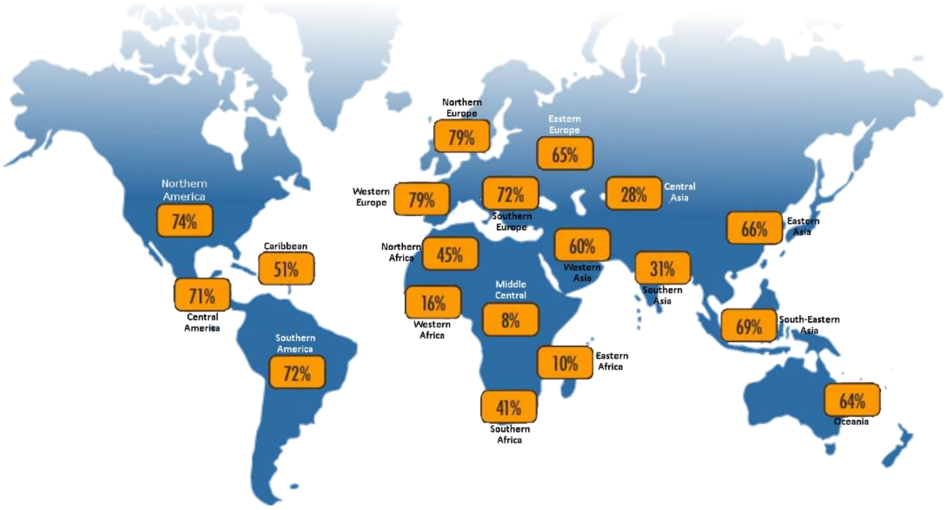


Figure 1 – Social media users by region vs. Total population by region (Kemp, 2021a)

In fact, social network platforms, commonly just referred to as social networks, have developed to be such an import global pivot point, that a new and distinct term, "social media," has been invented as a name for these software platforms, therefore, ensuring the same meaning across all areas.

With that said, while social media platforms strive for users to be able to generate content and engage with others, there are a variety of ways to do so (Voorveld et al., 2018). As a result, over the years, there have been and continue to exist various unique social media platforms, each with its own differentiating aspects. As such, this section focuses on presenting the considered social media platforms upon which the data will be collected for the sentiment and topic analysis, finishing with a comparison between the social medias with the most potential, along with a justification of the chosen option for the near real-time data extraction task of the proposed framework.

Each of the subsections of the considered social media platforms (Facebook, YouTube, Instagram, Twitter, Reddit, Tumblr, Twitch, and Discord) will present the results of the respective platform's study. Thus, each subsection will begin with the findings of the real-time availability of sports-related text content on that platform. If the findings are promising, information about the social media API will be provided, and only if the API is deemed adequate for the task will the demographic data of the users be presented.

2.1.1 Facebook

Facebook[5] is a social media, accessible through devices such as computers and smartphones, as presented in Figure 2, founded in 2004 and presently the most popular social network in terms of monthly active users (MAUs) (Kemp, 2021b), with 2.91 billion (Facebook, 2021a).

Figure 2 – Facebook feed. Retrieved from (Guynn, 2018)

This platform allows for the creation of profiles, pages, and groups, as means of presenting content and allowing users' interaction. Profiles are meant for users to present themselves to other users, either friends, if the profile is set to private, or strangers, when set to public. Pages are meant for brands, organizations, and public figures to engage with their fans or customers. And groups allow users to communicate about specific topics with other users who share common interests (Facebook, 2021b).

Although, currently the biggest social media platform, when it comes to the objective at hand, which is the near real-time collection of text content related to sporting events, Facebook has a few shortcomings that make it unsuitable as a data source. To begin with, data must be publicly available in order to be collected, which immediately excludes all content from private profiles. Secondly, despite the existence of sports-related groups, due to a lack of monitoring, the majority of them are filled with low-quality or unrelated content, often known as spam.

Finally, most posts on sports pages are done only in concern to specific important moments of the events, such as, in the case of football, goals and halftime, therefore their respective comments are likewise about that moment, however, usually, published with a significant delay

[5] Facebook Website - https://www.facebook.com/

to the time of the moment, thus, making it impossible to, through the posts and comments, create a timeline of the event.

2.1.2 YouTube

YouTube[6], founded in 2005 and accessible, among others, via computers and smartphones, as shown in Figure 3, is a social media dedicated to video sharing and livestreaming, being, currently, the second biggest social media with close to 2.3 billion MAUs (Kemp, 2021b).

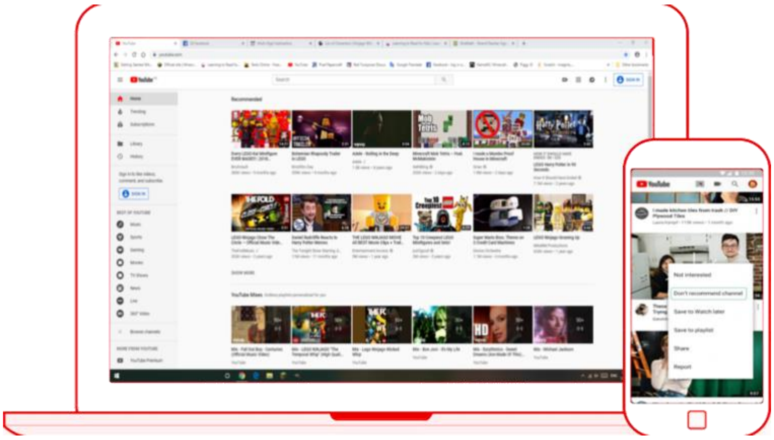
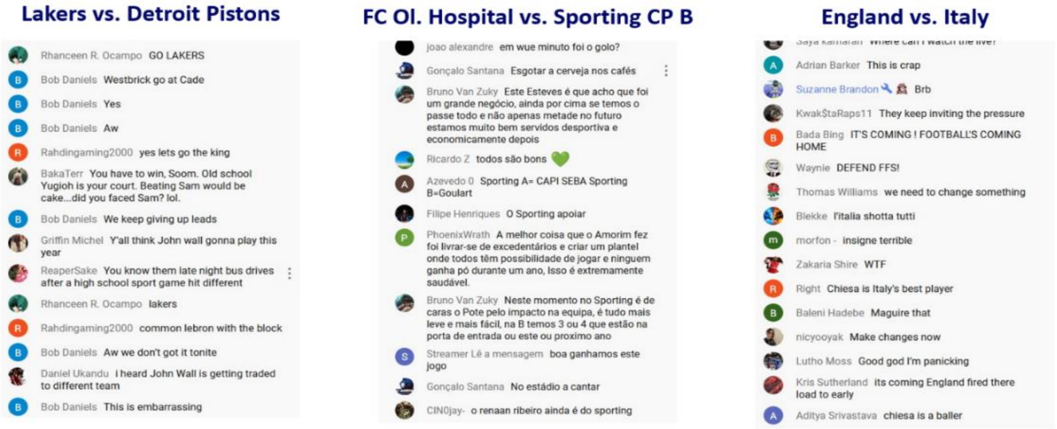


Figure 3 – YouTube feed. Adapted from (Guynn, 2018)

This platform, allows users to submit videos, perform livestreams, and comment on the available audiovisual content, therefore, enabling the understanding of the audience's perception of the content.

Although most known for its videos, YouTube, as previously noted, also enables its users to livestream, allowing its audience to comment in real time, via a live chat, on events as they unfold. Having said that, copyright regulations prevent any individual from just livestreaming a sporting event, with only a few events permitted to be livestreamed by official clubs or competitions' channels. Hence, the livestream's sports community has turned to reactions and watchalongs in order to overpass this limitation, enabling the YouTube catalogue of livestreamed sport events to be, despite the limitation, still rather extensive, including a wide range of games from different sports and competitions, as depicted by Figure 4.

[6] YouTube Website - https://www.youtube.com/



Lakers vs. Detroit Pistons

Rhanceen R. Ocampo GO LAKERS

Bob Daniels Westbrick go at Cade

Bob Daniels Yes

Bob Daniels Aw

Rahdingaming2000 yes lets go the king

BakaTerr You have to win, Soom. Old school Yugioh is your court. Beating Sam would be cake...did you faced Sam? lol.

Bob Daniels We keep giving up leads

Griffin Michel Y'all think John wall gonna play this year

ReaperSake You know them late night bus drives after a high school sport game hit different

Rhanceen R. Ocampo lakers

Rahdingaming2000 common lebron with the block

Bob Daniels Aw we don't got it tonite

Daniel Ukandu i heard John Wall is getting traded to different team

Bob Daniels This is embarrassing

FC Ol. Hospital vs. Sporting CP B

joao alexandre em wue minuto foi o golo?

Gonçalo Santana Esgotar a cerveja nos cafés

Bruno Van Zuky Este Esteves é que acho que foi um grande negócio, ainda por cima se temos o passe todo e não apenas metade no futuro estamos muito bem servidos desportiva e economicamente depois

Ricardo Z todos são bons 💚

Azevedo 0 Sporting A= CAPI SEBA Sporting B=Goulart

Filipe Henriques O Sporting apoiar

PhoenixWrath A melhor coisa que o Amorim fez foi livrar-se de excedentários e criar um plantel onde todos têm possibilidade de jogar e ninguem ganha pó durante um ano, Isso é extremamente saudável.

Bruno Van Zuky Neste momento no Sporting é de caras o Pote pelo impacto na equipa, é tudo mais leve e mais fácil, na B temos 3 ou 4 que estão na porta de entrada ou este ou proximo ano

Streamer Lê a mensagem boa ganhamos este jogo

Gonçalo Santana No estádio a cantar

CIN0jay- o renaan ribeiro ainda é do sporting

England vs. Italy

Saya Kamaran Where can I watch the live?

Adrian Barker This is crap

Suzanne Brandon Brb

Kwak$taRaps11 They keep inviting the pressure

Bada Bing IT'S COMING ! FOOTBALL'S COMING HOME

Waynie DEFEND FFS!

Thomas Williams we need to change something

Blekke l'italia shotta tutti

morfon - insigne terrible

Zakaria Shire WTF

Right Chiesa is Italy's best player

Baleni Hadebe Maguire that

nicyooyak Make changes now

Lutho Moss Good god I'm panicking

Kris Sutherland its coming England fired there load to early

Aditya Srivastava chiesa is a baller

Figure 4 – YouTube live chats of sports games' livestreams (EzBUCKETz, 2021; Mark Goldbridge That's Football, 2021; Sporting Clube de Portugal, 2021)

Regarding the extraction of comments from the live chats, YouTube provides an application programming interface (API), titled YouTube's LiveStreaming API[7], that works through a quota system. This API features a daily quota limit and a quota fee for each endpoint request, allowing requests to be made for free up to a limit of 10.000 units. With the GET endpoint for gathering live chat messages costing five units (Deliciousavocado, 2021), and providing up to 2000 comments per request, this API, therefore, allows the gathering of up to 4.000.000 comments per day. Additionally, YouTube allows for a paid daily limit increase, as well as the use of the API for commercial use, if its use complies with the YouTube API Services Terms of Service.

As for the demographics, YouTube, as presented in Figure 5, is mostly used by male users between 18 and 44 years old, with India and the United States of America having the most users, as shown in Figure 6.

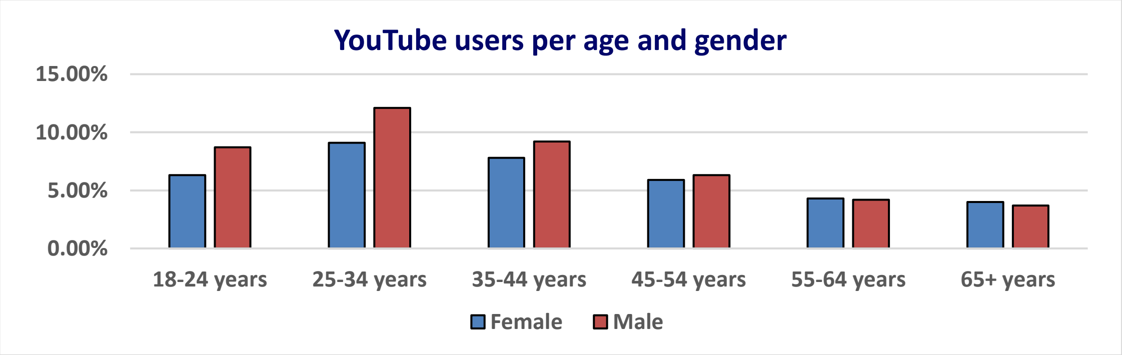


Figure 5 – YouTube users per age and gender (Kemp, 2021b)

[7] YouTube's LiveStreaming API Overview - https://developers.google.com/youtube/v3/live/

In despite of the number of users per country, regarding the number of visualizations per world region, Europe and North America are tied in second place, with Asia in first by only a slim margin, as also observed in Figure 6.

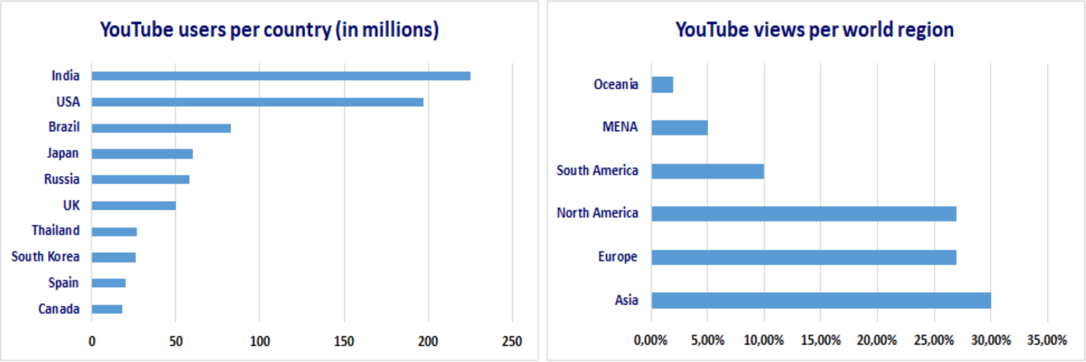


Figure 6 – YouTube users per country (GMI Blogger, 2021) and views per world region (ChannelMeter, 2019)

As a quick note, it is important to mention that in the graph relative to users by country, shown in Figure 6, all countries show their number of users, with exception of India, which shows the number of active users per hour, thus, having a much higher number of users.

2.1.3 Instagram

Instagram[8], originally developed exclusively for IOS, but now readily available through its Android app and browsers, as depicted in Figure 7, is a photo and video sharing social media founded in 2010, being, currently, the fourth biggest social media with close to 1.4 billion MAUs (Kemp, 2021b).

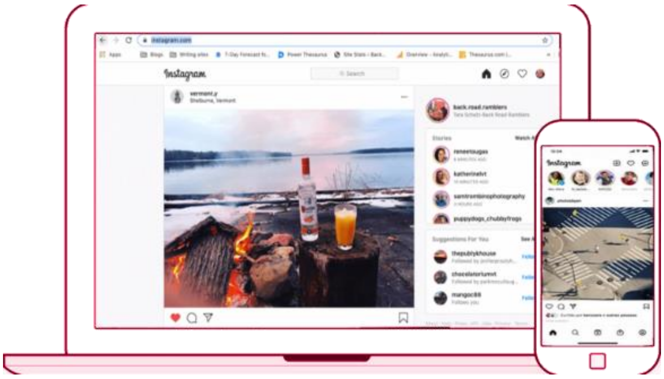


Figure 7 – Instagram feed. Adapted from (Guynn, 2018)

[8] Instagram Website - https://www.instagram.com/

This platform was created with the aim for its users to present themselves through photos, videos, and more recently, livestreams. Thus, despite allowing comments, due to its focus on audiovisual content, it has a very limited quantity of publicly available text content, therefore, excluding it from being a valid data source for the intended goal.

2.1.4 Twitter

Twitter[9], founded in 2006, is a social media focused on providing its users with interaction through means of microblogging, by only allowing messages up to 280 characters long (Gligorić et al., 2020).

This platform, currently, the fifteenth largest social media, due to its 436 million MAUs (Kemp, 2021b), and accessible through, both, computers and smartphones, as depicted in Figure 8, provides several tools as means of interacting with other users, namely, the ability to create posts, known as "tweets", comments and retweets, which consists in posts where a message can be added alongside an existing tweet.

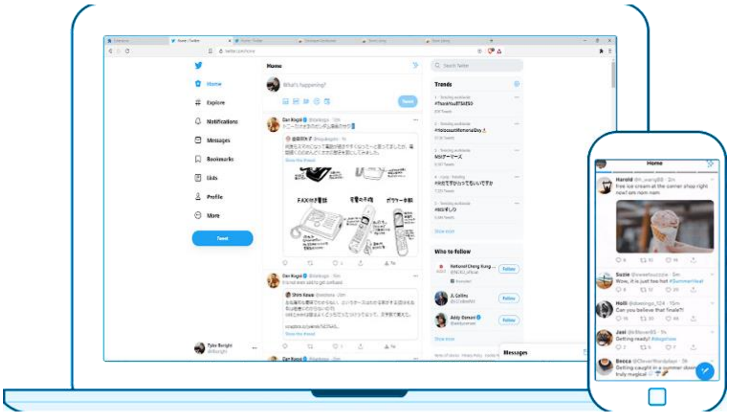


Figure 8 – Twitter feed. Adapted from (Guynn, 2018)

Nonetheless, it is the brief message format that most distinguishes this social media from others, leading to the added bonus of encouraging users to tweet more, as seen during the 2014 World Cup match between Brazil and Germany, where, during the respective telecast, 35.6 million match related tweets were published (Twitter Data, 2014). Furthermore, because of the characters restriction, messages are more packed with significant information when compared

[9] Twitter Website - https://twitter.com/

to most social media platforms, which, as a result, has made Twitter a heavily used tool and data source for a large number of studies, as seen, for example, in (Coelho, 2021).

Regarding the process of data collection, Twitter provides an API[10] with two versions, the v1 and v2, being the latter version more recent thus recommend, despite still being in early access. In addition, each API version has several access levels, thus allowing for a selection of features more appropriate for the intended use. Table 1 provides a recap of Twitter's v2 API access levels.

Table 1 – Twitter v2 API access levels (Twitter Developer Platform, 2021)

	Essential	**Elevated**	**Academic**
Cost	Free	Free	Free
Data volume	500.000 Tweets/month, 5 streaming rules, 512 characters	2.000.000 Tweets/month, 25 streaming rules, 512 characters	10.000.000 Tweets/month, 1000 streaming rules, 1024 characters
API v1 access	No	Yes	Yes
Full-archive access	No	No	Yes
Commercial use	Allowed	Allowed	Not allowed

In terms of demographics, as seen in Figure 9, Twitter is primarily made up of male users, with 75% of them being between the ages of 18 and 49.

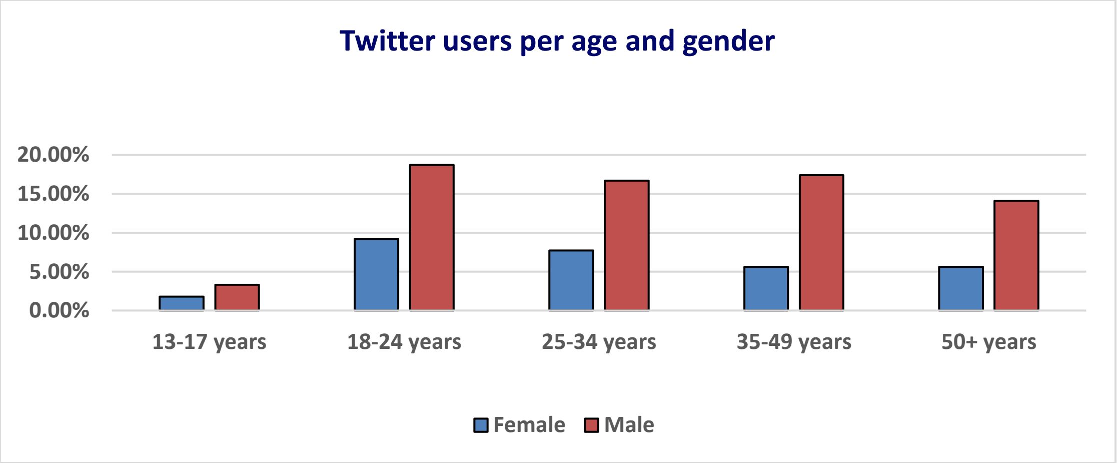


Figure 9 – Twitter users per age and gender (Kemp, 2021c)

As for the users per region, as shown in Figure 10, Asia dominates the platform having 44% of all global users, with North America and Europe coming, respectively, in second and third.

[10] Twitter API Overview - https://developer.twitter.com/en/products/twitter-api

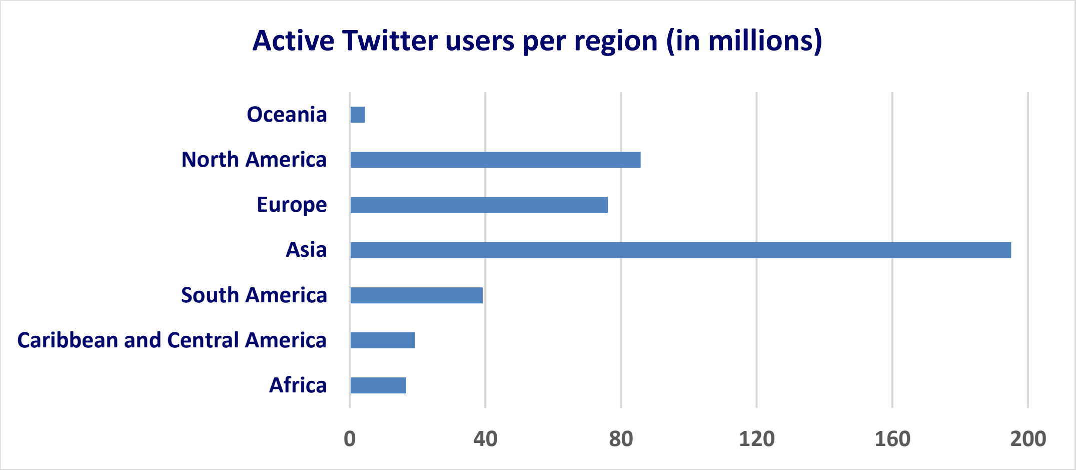
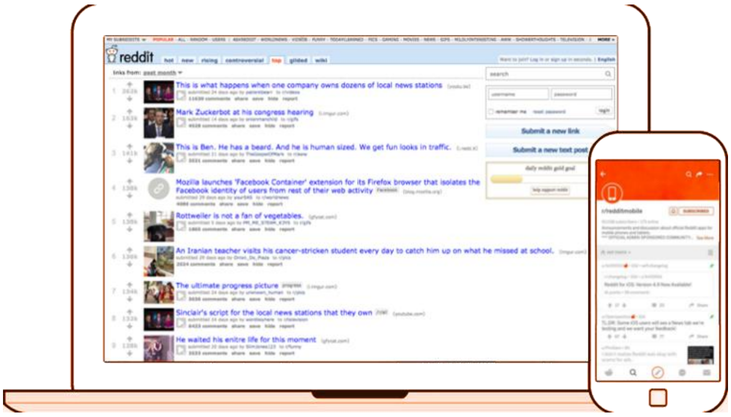


Figure 10 – Active Twitter users per region (Kemp, 2021c)

2.1.5 Reddit

Reddit[11] is a social media characterized by its forum-like format, having been founded in 2005 and accessible via, both, computers and smartphones, as shown in Figure 11, it is, currently, the sixteenth biggest social media, with 430 million MAUs (Kemp, 2021b).



This is what happens when one company owns dozens of local news stations

Mark Zuckerbot at his congress hearing

This is Ben. He has a beard. And he is human sized. We get fun looks in traffic.

Mozilla launches 'Facebook Container' extension for its Firefox browser that isolates the Facebook identity of users from rest of their web activity

Rottweiler is not a fan of vegetables.

An Iranian teacher visits his cancer-stricken student every day to catch him up on what he missed at school.

The ultimate progress picture

Sinclair's script for the local news stations that they own

He waited his enitre life for this moment

Figure 11 – Reddit feed. Adapted from (Guynn, 2018)

This platform allows users to submit posts on user-created communities, known as "subreddits," which are forums dedicated to specific themes, as well as comment on those posts, therefore, enabling information to be categorized by topics and, thus, making it easier to connect users with common interests.

[11] Reddit Website - https://www.reddit.com/

Furthermore, Reddit allows the use of bots to moderate and improve subreddits, providing them with the ability to automate certain actions. As such, some of the most popular sports-related subreddits have taken advantage of this, by automating the generation of match thread posts, whose main purpose is to provide its users with a place to discuss live sport events in real-time. Figure 12 provides an analysis done on the number of comments of 210 of these posts, being 70 in concern to the Champions League, 70 to the NBA, and 70 to the Premier League.

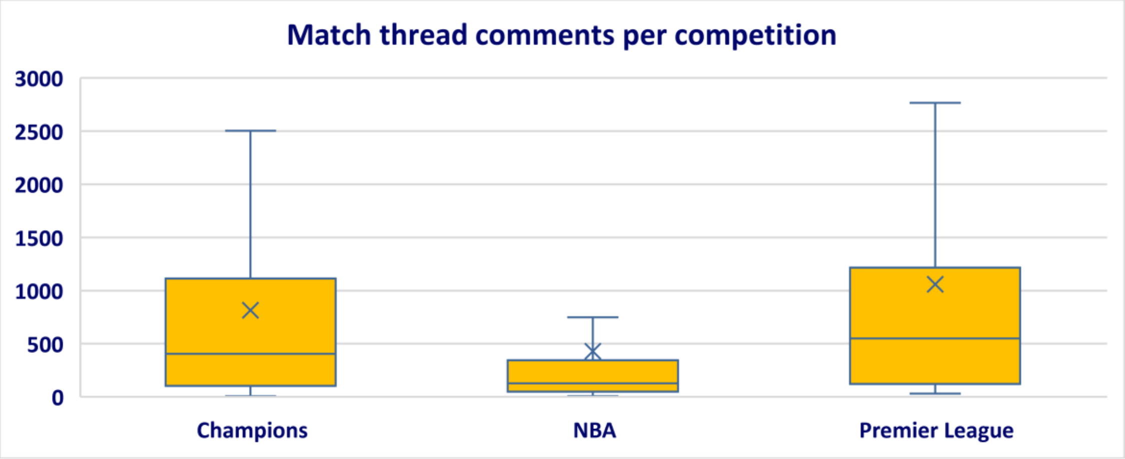


Figure 12 – Match thread comments per competition

As seen by the boxplots, after the removal of outliers, it is possible to conclude that, in Reddit, football live events are more discussed than basketball ones, despite, as shown in Figure 13, Reddit being dominated by United States users.

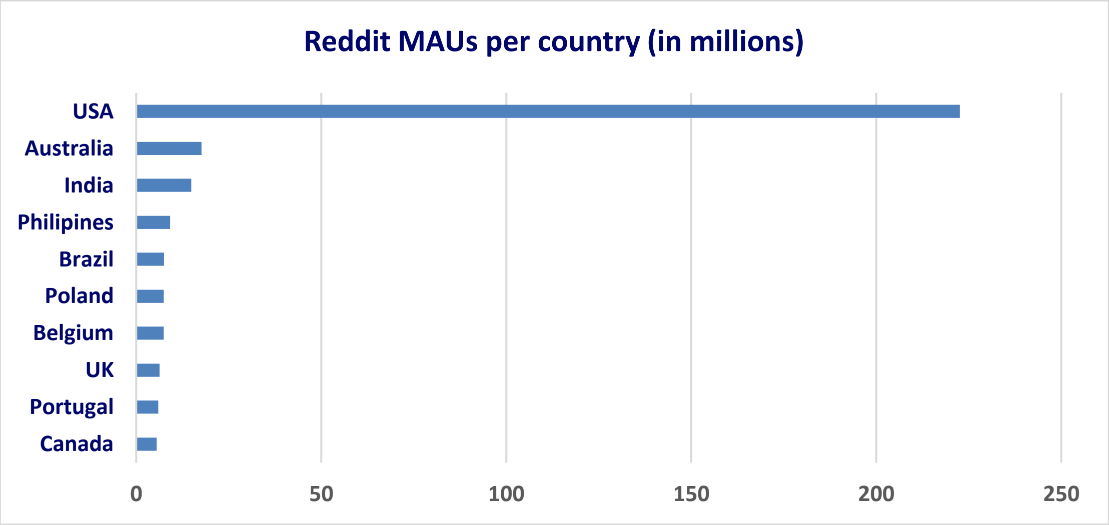


Figure 13 – Reddit MAUs per country (Degenhard, 2021)

Additionally, Reddit allows its subreddits to enable and manage users' flairs, which consist in tags that its users can choose to have connected with their accounts. This has enabled sports-related subreddits to create flairs for most first league teams and national teams, thus enabling comments to have as context the writer's supporting team.

Regarding the data collection process, Reddit provides a free API[12], which, upon Reddit's approval, can be used for commercial purposes. This API allows for up to 60 requests per minute, with the comments' retrieval GET endpoint gathering up to 100 comments per request, thus enabling a collection of up to 8.640.000 comments per day. With this said, a third-party wrapper, named Async PRAW[13], has been developed in order to, taking advantage of the high-rate limit of the API, provide the API users with a streaming option for the gathering of the data in real-time.

In terms of demographics, Reddit is dominated by male users, with only 37.2 percent of users being female (Kemp, 2021b). As for ages, most users, around 54%, are between 20 and 40 years old, as presented in Figure 14.

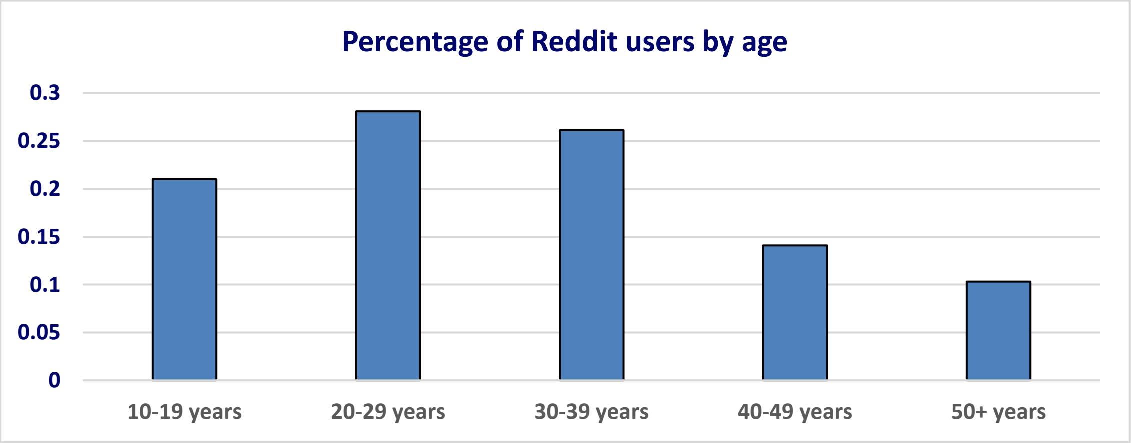


Figure 14 – Percentage of Reddit users by age (Statista, 2021)

2.1.6 Tumblr

Tumblr[14] , founded in 2007 and accessible by both computer and smartphone, as shown in Figure 15, is a social media focused on providing its users with blogs, instead of profiles.

[12] Reddit API Overview - https://www.reddit.com/dev/api/

[13] Async PRAW Overview - https://asyncpraw.readthedocs.io/en/stable/

[14] Tumblr Website - https://www.tumblr.com/

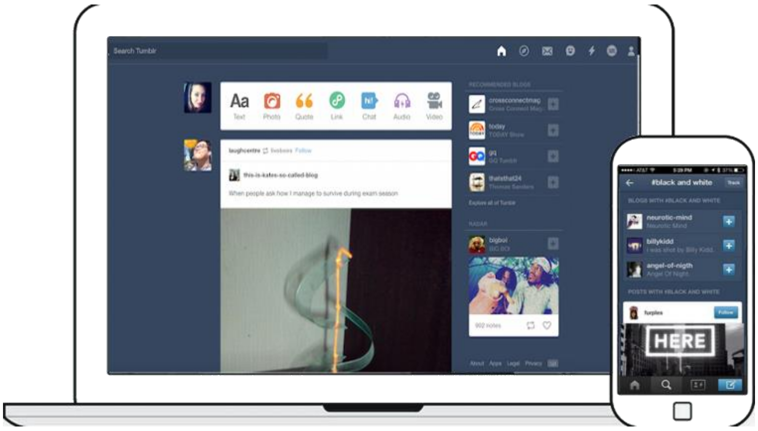


Figure 15 – Tumblr feed. Adapted from (Guynn, 2018)

This platform does not promote the writing of comments, only allowing them to be added through the reblog of existing posts. This, in turn, leads to a very small number of comments, making this social media unsuitable for the collection of text content.

2.1.7 Twitch

Twitch[15], accessible by, both, computers and smartphones, as depicted in Figure 16, is a social media centered around livestreaming that was founded in 2011, having, currently, 140 million MAUs (Dean, 2021).

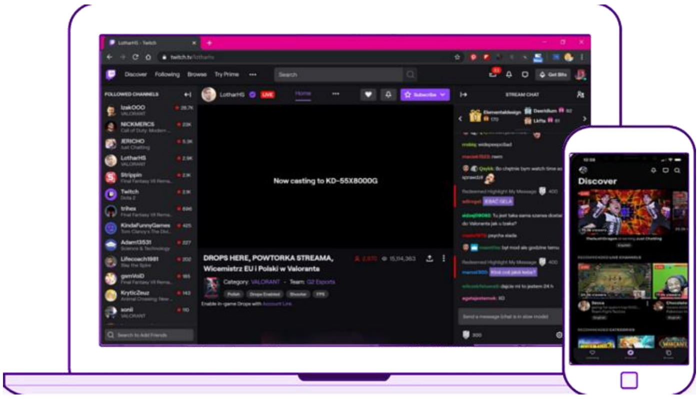


Figure 16 – Twitch feed. Adapted from (Guynn, 2018)

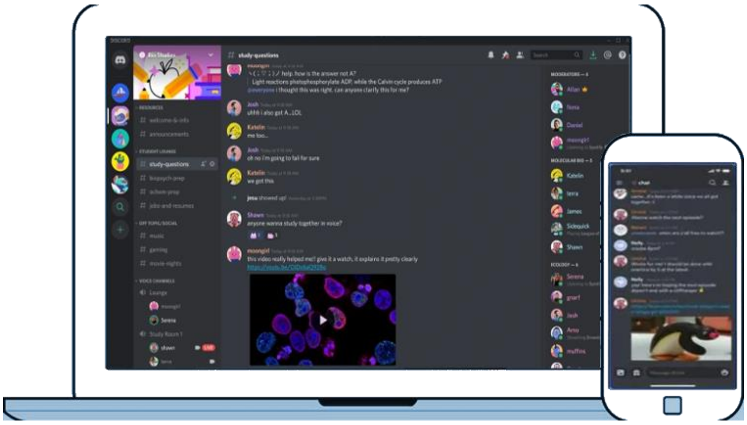
This platform allows users to livestream or watch other users' livestreams, as well as comment in real time, via live chat, on events as they happen. Although similar to YouTube, due to this platform's focus on video gaming-related material, as well as copyright regulations, the number

[15] Twitch Website - https://www.twitch.tv/

of sports-related livestreams, particularly of real-time events, is extremely limited, excluding this social media from being a reliable data source for the intended objective.

2.1.8 Discord

Discord[16], founded in 2015 and available for both computers and smartphones, as presented in Figure 17, is an instant messaging and Voice over Internet Protocol (VoIP) social media, with 140 million MAUs (Curry, 2021).



moongirl
help. how is the answer not A?
Light reactions photophosphorylate ADP, while the Calvin cycle produces ATP
@everyone i thought this was right. can anyone clarify this for me?

Josh
uhhh i also got A...LOL

Katelin
me too...

Josh
oh no i'm going to fail for sure

Katelin
we got this

jesu showed up!

Shawn
anyone wanna study together in voice?

moongirl
this video really helped me!! give it a watch, it explains it pretty clearly

Figure 17 – Discord feed. Adapted from (Guynn, 2018)

This platform allows its users to interact with each other using text messaging, voice calls and video calls, through either, private chats or servers, which are an invitation-only user-made collection of persistent chat and voice rooms.

With that said, although having servers dedicated to the real-time discussion of sporting events, this social media is not ideal as a data source for the intended framework, as only users with Administrative or "Manage Server" permissions on each server can add bots, which are the only way available for the real-time gathering of server messages. As a result, for the addition of a bot into a server, it is necessary to request to the server staff members, with the required permissions, to add the bot, which does not ensure that they will do so, especially when the bot's only role is to gather messages, thus not improving the server in any manner.

[16] Discord Website - https://discord.com/

2.1.9 Social Networks' comparison

Considering the study results of the aforementioned social media platforms, it is possible to conclude that of the eight presented platforms, YouTube, Twitter and Reddit are the ones with the most potential for the real-time, or near real-time, gathering of sport events text content.

As such, Table 2 exposes a comparison between the three social medias, with each line having a different characteristic deemed of importance for the intended purpose. The aspects compared were cost, data volume, request type, commercial use, content, and sports events availability.

Table 2 – Comparison between YouTube, Twitter, and Reddit

	YouTube	**Twitter**	**Reddit**
Cost	Free	Free	Free
Data volume	4.000.000 comments/day	2.000.000 tweets/month	8.640.000 comments/day
Request type	GET	Streaming	Streaming
Commercial use	Allowed	Allowed	Allowed, upon Reddit's approval
Content	Quality and quantity vary substantially	Good quality and quantity	Good quality, but quantity varies substantially
Sports events availability	Most from major leagues and competitions. And some from minor leagues and competitions	All or close to all	Most from major leagues and competitions.

Having concluded the comparison between the three mentioned platforms, from the author's perspective, Twitter is the best suited social media for the real-time gathering of sports related text content.

This decision was based on the better availability of sports events discussion, typically higher content quality and quantity, and the ability to stream the data rather than just provide it via GET requests. As for the data volume, although Twitter's API allows for a considerably lower amount of data gathering per month than YouTube and Reddit, due to the proposed framework support for multiple data sources, this aspect importance is reduced, as the long-term aim is for the data to be collected from multiple social medias, thus complementing this weak data volume. At last, despite the predominance of male users, the more representative diversity of ages among Twitter users also contributed to this choice.

2.2 Text Mining

Humans, by nature, are insatiably curious, being always eager to learn new things. As such, from the dawn of modern computers, researchers have idealized the use of these machines as a tool that can, not only assist people in the process of knowledge discovery but also learn new knowledge or information on its own (Turing, 1950).

As a result, since its inception, this area of study, titled machine learning (ML) and encompassed within the AI field, has provided various unique techniques for the self-discovery of knowledge. However, it was in part due to the creation of the relational database, which allowed for an efficient storing and gathering of data (Codd, 1970), that this field started to evolve at a better level, thus leading to the coining of the term knowledge discovery in databases (KDD), which according to (Fayyad et al., 1996) is a “non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data”.

KDD, however, is not a technique to extract knowledge from data, but rather a series of steps that make up the data science life cycle (Hotho et al., 2005), with the process of extracting information from the data, referred to as data mining, accounting for only a portion of the process. As such, KDD provides for the ability to select the pre-processing and processing techniques of choice based on the data and task at hand, being focused on analysing structured data (Fayyad et al., 1996), which is data that follows a standardized format. With this said, due to most publicly available online data being in the format of text documents, thus unstructured (IBM, 2021), it was necessary for the introduction of a process for knowledge discovery in text (KDT) (Feldman & Dagan, 1995).

KDT, similar to KDD, is a set of steps for the data science life cycle, however focused on extracting information from text. As such, this process centred around NLP, allows for the selection of the text mining, a data mining subfield, technique of choice, enabling the analysis of text for a variety of tasks, such as text categorization, text clustering, sentiment analysis, document summarization, topic modelling, among others (Aggarwal & Zhai, 2013; Han et al., 2012).

2.2.1 Sentiment Analysis

Human to human communication is a complex process that involves both verbal and non-verbal actions. As such, in order to accurately understand a verbal message, one must, not only understand the language, but also the underlying signals given by the person speaking, usually noticeable through the body language (DeVito et al., 2000). To simulate the same behaviour in a human-machine interface it is, therefore, essential that computers can detect the human affective state (Sterley & Bains, 2021). Thus, as a result, the field of affective computing was introduced with the aim of enabling computers with techniques for the recognition, interpretation and simulation of human emotions (Picard et al., 2004).

Human communications, however, can present themselves in a written format, as such the understanding of emotion in text began with studies aimed at understanding how text could express or generate different emotions (An et al., 2017; Daily et al., 2017; Lutz & White, 1986; Osgood et al., 1975). These studies, however, did not focus on the computer aspect of affective detection, but rather on the human one, as such the initial techniques for sentiment analysis in text were centred around the sentiment of each word (Taboada et al., 2011), aiming to classify the text polarity as, either, positive, negative or neutral.

These techniques based on the lexicon, however, had several limitations, as such, as computers became more powerful, a new type of techniques based on ML appeared (Medhat et al., 2014). Figure 18 illustrates some of the lexicon and ML based most used techniques for sentiment analysis.

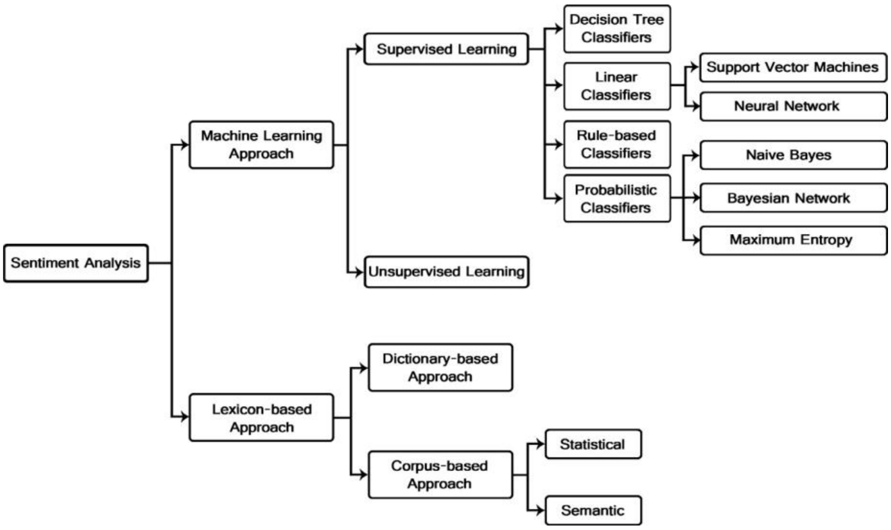


Figure 18 – Sentiment analysis most used techniques (Medhat et al., 2014)

2.2.1.1 Lexicon Approaches

Lexicon-based sentiment analysis works by calculating the text document's polarity from the semantic orientation of lexicons (N. Gupta & Agrawal, 2020), either through a dictionary-based or corpus-based technique.

Dictionary-based or rule-based approaches

The sentiment analysis procedure in dictionary-based approaches, as the name implies, relies on dictionaries to determine the polarity of words, thereby, working by gathering the sentiment values of all dictionary words included in the text document, in order to calculate its sentiment.

As such, these approaches are highly dependent on the dictionary used, reason why, over the years, there have been several iterations of dictionaries, each, aiming to improve on the existing ones. The General Inquirer (GI[17]) was one of the initial works on this field (Hartman et al., 1967). This application for content analysis, provided over 11000 words categorized into one, or more, of 183 categories, including over 4000 words labelled as, either, positive or negative (Hutto & Gilbert, 2014). Similarly, the Linguistic Inquiry and Word Count (LIWC[18]) (Tausczik & Pennebaker, 2010), developed for the measuring of thoughts, feelings, personality and motivations, contained 4500 words categorized into one, or more, of 76 categories, of which, 905 words were, either, classified as positive or negative (Hutto & Gilbert, 2014).

These early works, however, were limited by their binary classification of emotion, therefore making all words from a category have the same connotation. As a result, a new type of dictionary was introduced, in which words, rather than labelled, were associated with a valence score for sentiment intensity (Hutto & Gilbert, 2014). The Affective Norms for English Words (ANEW[19]), thus, was created, offering a ranking of its words according to their pleasure, arousal, and dominance (Bradley & Lang, 2008). SentiWordNet[20], which is an extension of WordNet[21], an English lexical database (Miller, 1995), was also created for the ranking of words, however, instead of ranking single words relative to their positivity, negativity, and objectivity, it focused on ranking synsets (Esuli & Sebastiani, 2006), which are groups of synonyms that express the same concept.

[17] GI Website - http://www.wjh.harvard.edu/~inquirer

[18] LIWC Website - http://liwc.wpengine.com/

[19] ANEW Website - https://csea.phhp.ufl.edu/media/anewmessage.html

[20] SentiWordNet Repository - https://github.com/aesuli/SentiWordNet

[21] WordNet Website - https://wordnet.princeton.edu/

These new approaches, although an improvement from prior approaches, were still flawed as they did not consider the context of the words, as such, VADER[22] was introduced (Hutto & Gilbert, 2014). VADER, meaning Valence Aware Dictionary and sEntiment Reasoner, is a rule-based model for sentiment analysis, which additionally from using a dictionary for the gathering of the valence score of the words, introduced the use of word-sense disambiguation (Akkaya et al., 2009) as means of understanding the words' context (Hutto & Gilbert, 2014). Additionally, having been developed with social media sentiment analysis in mind, this technique, also improved on prior ones, by adding emoticons to its dictionary (Ayvaz & Shiha, 2017), thus proving to be one of the better lexicon approaches by being able to provide fast, but relatively accurate sentiment classifications (Bonta et al., 2019; Hutto & Gilbert, 2014).

Corpus-based approaches

Sentiment analysis based on corpus relies on co-occurrence statistics or syntactic patterns embedded in text corpora (Darwich et al., 2019), which are large, structured sets of texts. Thus, the idea behind these approaches is that the distance between a word and a set of positive and negative seed words (Jovanoski et al., 2016) can be used as a metric to estimate its polarity (Darwich et al., 2019). This distance, however, can, either refer to the distance in the text between words, or the semantical similarity between words. As such, inside the sentiment analysis corpus-based approaches there are two types of techniques. The ones based on statistics and the ones based on semantic (Rajput & Solanki, 2016).

Statistical corpus-based approaches work by estimating words' polarities by calculating their relative frequency of co-occurrence with another words (Rajput & Solanki, 2016). As seen in (Velikovich et al., 2010), where it was proposed a graph propagation model, constructed from co-occurrence statistics from the entire web, to derive a polarity lexicon through the highest weight between seed nodes and target nodes. Besides label propagation (Huang et al., 2014), pointwise mutual information (PMI) can also be used as a statistical approach to sentiment analysis, as proposed in (Turney, 2002), where the semantic orientation can be calculated as the PMI between the given phrase and the word "excellent", minus the PMI between the given phrase and the word "poor".

[22] VADER Repository - https://github.com/cjhutto/vaderSentiment

Semantic corpus-based approaches, on the other hand, work by following the principle that semantically close words have similar sentiment values (Rajput & Solanki, 2016). As such, these type of approaches take advantage of semantic models, such as WordNet (Miller, 1995), to classify word sentiment opinions based on their synonyms and antonyms (Araque et al., 2019). Another approach is presented in (Kim & Hovy, 2004), where it is proposed the use of the relative count of positive and negative synonyms as an estimation of words' sentiments.

2.2.1.2 Machine Learning approaches

ML-based sentiment analysis works through the inference of text's semantic orientation by models trained on text data. As a result, depending on the data labelling state, these approaches can be divided into either: supervised, unsupervised, or semi-supervised techniques.

Supervised techniques

Supervised ML consists in the training of models using labelled data points. As such, for the task of sentiment analysis using supervised learning algorithms, it is imperative that the data used consists in text documents labelled with their polarity, thus, severely limiting the availability of useful datasets.

Regarding the choice of technique the number of available options is not as limited, therefore, for the task of sentiment analysis there have been used a wide range of algorithms, such as the ones of Maximum Entropy (H. Lee & Bhd, 2011), Bayesian Networks (Gutiérrez et al., 2019), Naïve Bayes (Parveen & Pandey, 2017), Support Vector Machines (SVM) (Ahmad, Aftab, & Ali, 2017), Neural Networks (NN) (Tul et al., 2017), among others (Ahmad, Aftab, Muhammad, et al., 2017).

Of the aforementioned algorithms, NN, by norm, achieve the better results (Abd El-Jawad et al., 2019). Therefore, the sentiment analysis SOTA study presented in this subsection focuses on further exposing techniques based on this algorithm, which works by replicating the functioning of the human brain through the use of nodes, to represent neurons, and edges, to represent synapses, which are the connections between neurons (Asadollahfardi, 2015).

This, apparently simple, structure of NN, in turn, allows for a high level of customization, enabling the creation of NN, whose node types and architecture are adapted to the task at hand. Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are two of the most popular types of NN, being, respectively, good at extracting position-invariant features and modelling units in sequence (Yin et al., 2017), thus making CNN primarily used to deal with images and RNN with text. Nonetheless, as presented in (Yin et al., 2017), CNN can also be used for text classification tasks, being able to achieve performances comparable to RNN.

RNN, however, although having been developed with text in mind, suffer in performance when used with longer text sequences. As such, in 1997 the Long Short-Term Memory (LSTM) architecture was introduced as a way to deal with this limitation (Hochreiter & Schmidhuber, 1997). Despite achieving better performances, LSTM are computationally expensive to train as they are not designed to be parallelizable. So, as a result, with the aim of solving this limitation and further improving performances for longer text sequences, (Vaswani et al., 2017) proposed the Transformer, an encoder-decoder-based neural network with an attention mechanism, which, by dispensing the need for recurrence and convolutions, reduces significantly the training time, while achieving SOTA results.

This new architecture proved to be revolutionary for the NLP field, thus becoming a steppingstone for many new language representation models, such as Bert (Devlin et al., 2019), which stands for Bidirectional Encoder Representations from Transformer and works by stacking several Transformer' encoders. This model, pre-trained with 800 million words from the BooksCorpus and 2500 million words from the English Wikipedia, allows for the fine-tuning for several NLP tasks, having achieved SOTA results in eleven of them. According to (Y. Liu et al., 2019), BERT, however, is significantly undertrained, thus, this paper introduces a new model named RoBERTa, which, besides having been trained with more data, removes the Next Sentence Prediction (NSP) task from the pre-training and introduces dynamic masking, therefore, achieving SOTA results on GLUE (A. Wang et al., 2019), RACE (Lai et al., 2017) and SQuAD (Rajpurkar et al., 2016).

Large pre-trained language models (PLM), however, due to their large dimensions are very computational expensive, and thus slow, to train and use for inference. As such, in (Sanh et al., 2019), it is proposed a smaller general-purpose language representation model based on BERT, titled DistilBERT. This model, using knowledge distillation, which is the process of transferring knowledge from a larger model into a smaller one, is able to achieve a reduction in size of 40% and an improvement in speed of 60%, while retaining 97% of the language understanding capabilities. Following the same logic, knowledge distillation was also performed on the RoBERTa model for the creation of the DistilRoBERTa model (Hugging Face, 2019).

Transfer learning, and consequently pre-trained models, have changed the paradigm of NLP, however extreme fine-tuning of this models can lead them to overfit and forget the pre-trained knowledge. As such, (Jiang et al., 2020) proposes a framework for the fine-tuning of pre-trained language models, which, through smoothness-inducing regularization and Bregman proximal point optimization, aims to solve the aforementioned issues. This framework is able to achieve SOTA results on multiple NLP benchmarks, such as the one of sentiment analysis in the SST-2 Binary classification dataset.

Although PLM, as already mentioned, are able to be fine-tuned, not all domain knowledge can be represented in a way useful for the fine-tuning process. As such, (Ostendorff et al., 2019) proposes a knowledge embedding process, which allows for the enriching of the BERT model with metadata and knowledge graph embeddings, thus achieving better results in a books' classification task than standard BERT. (W. Liu et al., 2020) also proposes the use of knowledge graphs as a way of embedding knowledge into BERT, by allowing the injection of triples into sentences. However, as embedding to much knowledge can divert the sentence meaning, this approach also introduces the use of soft-position and visible matrix as ways of limiting the impact of knowledge. At last, (X. Wang et al., 2021) also proposes a technique for embedding knowledge into PLM. Similarly to the others, this approach uses knowledge graphs, however, instead of resorting to the graph entities embeddings, it uses the textual entities description embeddings for the joint optimization with the PLM embeddings.

Previous approaches assume that for the task at hand, one has sufficient data for the fine-tuning. However, that may not be the case, as such, in (S. Wang et al., 2021), a new approach for fine-tuning, titled EFL, which stands for Entailment as Few-Shot Learner, is proposed. The key idea behind EFL consists in reformulating the potential NLP task into an entailment one, thus enabling the fine-tuning of a PLM with, as little, as 8 examples. This approach, which can be combined with an unsupervised contrastive learning-based data augmentation method, is able to achieve better SOTA results when compared to other few-shot learning methods, being able to yield competitive few-shot results with much larger models, such as GPT-3 (Brown et al., 2020).

Unsupervised techniques

Unsupervised ML, contrary to supervised ML, works by allowing the identification of patterns from unlabeled datasets.

As such, in (Ma et al., 2017) a comparative study on clustering-based sentiment analysis was done. This study concluded that clustering algorithms of the K-Means' type performed better on balanced datasets, with K-Means (Mannor et al., 2011) achieving the highest average accuracy of all clustering algorithms studied. Regarding unbalanced datasets, Agglo-WSlink (Zhao & Karypis, 2005), Slink (Jain et al., 1999), UPGMA (Kaufman & Rousseeuw, 1990), Spect-Sy (Ng et al., 2002), Spect-RW (Shi & Malik, 2000), PCA-Kmeans (Pearson, 1901) and Spect-Un (Von Luxburg, 2007) proved to be the best suited clustering algorithms.

In (Radford et al., 2017), while training a large multiplicative LSTM (Krause et al., 2019) on next character prediction of Amazon reviews, it was discovered a sentiment neuron able to highly predict sentiment values. Although, not intentionally, this paper, therefore, demonstrates that through the training of large NNs in fake tasks, it may be possible to learn accurate sentiment analysis classification.

Semi-supervised techniques

Semi-supervised techniques are meant to be used in situations where one has a dataset where only a small portion of the instances are labelled. This type of techniques, thus are a hybrid approach between supervised and unsupervised techniques.

In (Xie et al., 2020) it is proposed a data augmentation technique, which, using advanced data augmentation methods, such as RandAugment and back-translation, is able to achieve SOTA results for the IMDB text classification dataset, with only 20 labeled examples, thus beating the previous SOTA model that had been trained on 25000 examples.

2.2.2 Topic Analysis

Topic analysis, otherwise known as topic modelling or topic mining, consists in a "technique for revealing the underlying semantic structure" (Kherwa & Bansal, 2018) of a collection of documents, thereby, enabling the discovery of its topics and, consequently, allowing the classification of new documents according to the learnt topics.

As such, when performed with topic inference in mind rather than keyword discovery, topic analysis, like sentiment analysis, can be considered a classification task. However, unlike sentiment analysis, where labels are normally consistent across datasets (usually positive, neutral, and negative), topic analysis labels vary widely, making it increasingly difficult to discover suitable labelled datasets for the task at hand. As a result, unsupervised statistical techniques have dominated this field of NLP (Kherwa & Bansal, 2018).

One of these techniques is the Non-Negative Matrix Factorization (NMF) (D. D. Lee & Seung, 1999). This statistical method works by reducing the dimension of the text corpora, through the use of the factor analysis method, thus assigning less weight to less coherent words. In (Yan et al., 2013), it is proposed a topic modelling approach for short texts. This approach works by applying symmetric NMF on the term correlation matrix, thus, instead of using high-dimensional and sparse term occurrence data, it uses correlation data. In (Dumais, 2004) it is proposed the use of Latent Semantic Analysis (LSA) as an approach to topic modelling. LSA works by using singular value decomposition (SVD) to analyze the relationships between documents and their words. Thus, this technique consists in the decomposition of a matrix of documents and terms into a document-topic matrix and a topic-term matrix.

Another approach consists in the Latent Dirichlet Allocation (LDA) (Blei et al., 2003), which is a generative probabilistic model able to classify text documents relative to their topics. This Bayesian model works through a three-level hierarchical structure, where each document can be described by a distribution of topics and each topic by a distribution of words. In (Moody, 2016), LDA is combined with the skip-gram architecture of word2vec (Mikolov et al., 2013) to create lda2vec. This new technique works by learning word vectors to obtain sparser topic vectors that are easier to interpret.

The preceding techniques, however, are not designed with near real-time topics' classification in mind. As such, in (Yao et al., 2009) it is presented a sampling technique, which combined with the use of LDA aims to improve the topics identification computation time. For this, three sampling techniques are presented. Gibbs1, which jointly resamples all topics, Gibbs2, which jointly resamples topics for all new documents, and Gibbs3, which independently resamples topics for new documents, thus allowing for the processing of all documents independently. In (Yuan et al., 2015) a new model is proposed with the intention of lowering the computational expense often associated with topic modelling while simultaneously cutting computation times.

This model, named LightLDA, improves on existing models and techniques by combining the Metropolis-Hastings sampling algorithm, which has a running cost independent of model size, with a model storage data structure that uses separate data structures for high and low frequency words, allowing large models to fit in memory while maintaining high inference speeds. Additionally, this model uses a bounded asynchronous data-parallel technique that allows for distributed processing, enabling the training of a 1 trillion parameter topic model on as few as eight machines.

The aforementioned techniques, however, do not take into consideration the relationships between topics, which can be represented trough a hierarchy structure. Furthermore, due to their completely unsupervised character, none of the above-mentioned techniques consider the desired topics, thus usually deviating from them. As such, (Meng et al., 2020) proposes a joint spherical space embedding topic mining model, named JoSH, which works by using directional similarity to characterize semantic correlations among words, documents and categories, while allowing for the use of a category tree, described by category names, as a way to guide the learning process. Another weakly-supervised hierarchical technique, presented in (Meng et al., 2019) and named WeSHClass, proposes a neural approach, which using weak supervision through the form of class-related documents or keywords generates pseudo-documents to pretrain the model.

At last, most text modelling techniques work only by presenting the results, thus failing in providing the uncertainty of those same results. As such, (Kesiraju et al., 2020) proposes a SOTA Bayesian subspace multinomial model that learns to represent documents in the form of Gaussian distributions, thereby encoding the uncertainty in its co-variance.

2.3 Text Mining applied to Social Networks

After an extensive market research, it was concluded that currently there are several other tools, whose purpose is aligned with the one of the data fusion and analysis framework. As such, the following subsections present tools able to collect and analyze tweets in regard to their sentiments, topics, or both.

2.3.1 Twitter Sentiment Visualization

Twitter Sentiment Visualization[23], shown in Figure 19, is a free web-based tool developed for the gathering of tweets, through Twitter's search API, and, consequently, analysis and visualization of their sentiments, with confidence measurement of the estimates (Healey & Ramaswamy, 2013).

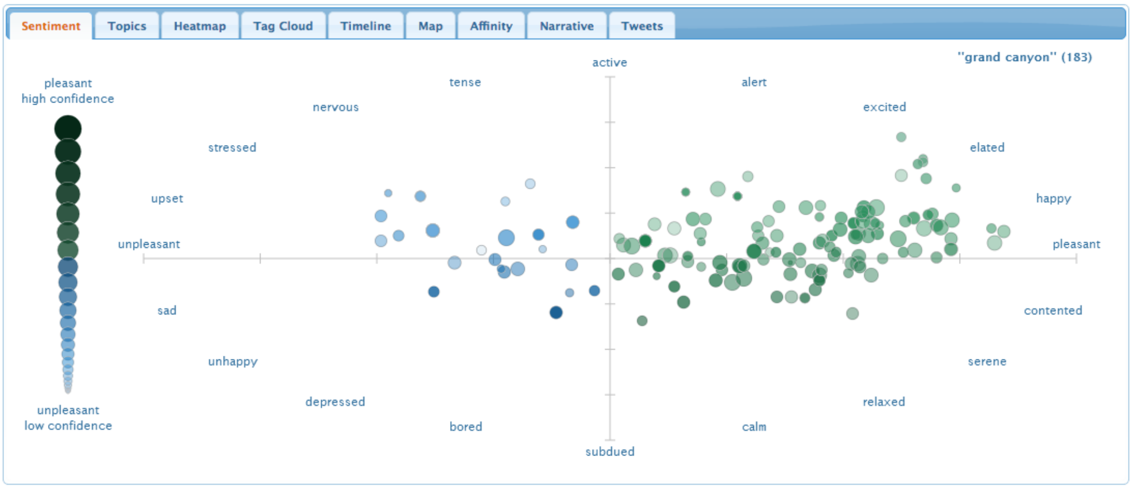


Figure 19 – Twitter Sentiment Visualization Dashboard (Healey & Ramaswamy, 2013)

This tool works by using a lexicon-based approach for the sentiment analysis, using a combination of the extended ANEW (Warriner et al., 2013) and happiness dictionaries (Dodds et al., 2011), where each word has a mean rating and a standard deviation for, both, a measure of valence and arousal (Healey & Ramaswamy, 2013). As such, the sentiment estimative is done by performing a weighted average between the words of the tweet, which are in the dictionaries, being the weights the results of the probability density function of a normal distribution of each of the words (Healey & Ramaswamy, 2013).

Being centred around providing the tweets' analysis visually, this tool provides its users with the following visualization tabs:

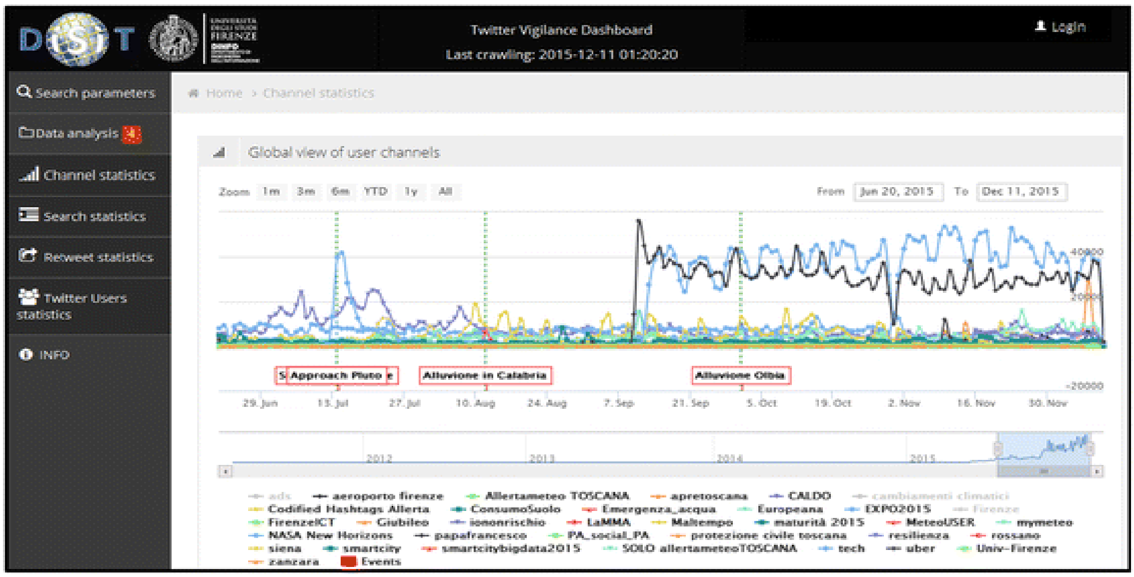
- **Sentiment tab**: Tweets are presented in a scatterplot, where the horizontal and vertical axes correspond, respectively, to pleasure and arousal levels.
- **Topics tab:** Tweets are clustered according to their text similarity.
- **Heatmap tab:** Shows an heatmap of the sentiment scatterplot.

[23] Twitter Sentiment Visualization Dashboard - https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/

- **Tag cloud tab:** Shows the most frequently occurring terms of four emotional regions (upset, happy, relaxed, and unhappy).
- **Timeline tab:** Bar graph that shows the timeline of the tweets, along with their emotional region.
- **Map tab:** Displays tweets' posting locations on top of a world map.
- **Affinity tab:** Provides a graph visualization of the relationships (or affinities) between tweets, people, hashtags, and URLs.
- **Narrative tab:** Presents narrative threads, which are sets of tweets that form conversations about a common topic over time.
- **Tweets tab:** Presents a table, with the collected tweets' dates, authors, bodies, and overall pleasure and arousal values.

2.3.2 Twitter Vigilance

Twitter Vigilance[24], shown in Figure 20, is a multi-user analysis tool, capable of monitoring, both, slow and fast Twitter events through a Crawler, which is a multithread system that exploits the Twitter Search API (Crisci et al., 2018). Thus allowing for the sentiment analysis, through a lexicon-based approach, which computes the sentiment based on the tweets' extracted adjectives, of 98% of the tweets and retweets collected from the monitored events.



Twitter Vigilance Dashboard
Last crawling: 2015-12-11 01:20:20

From Jun 20, 2015 To Dec 11, 2015

Figure 20 – Twitter Vigilance Dashboard (Crisci et al., 2018)

[24] Twitter Vigilance Dashboard - http://www.disit.org/tv/

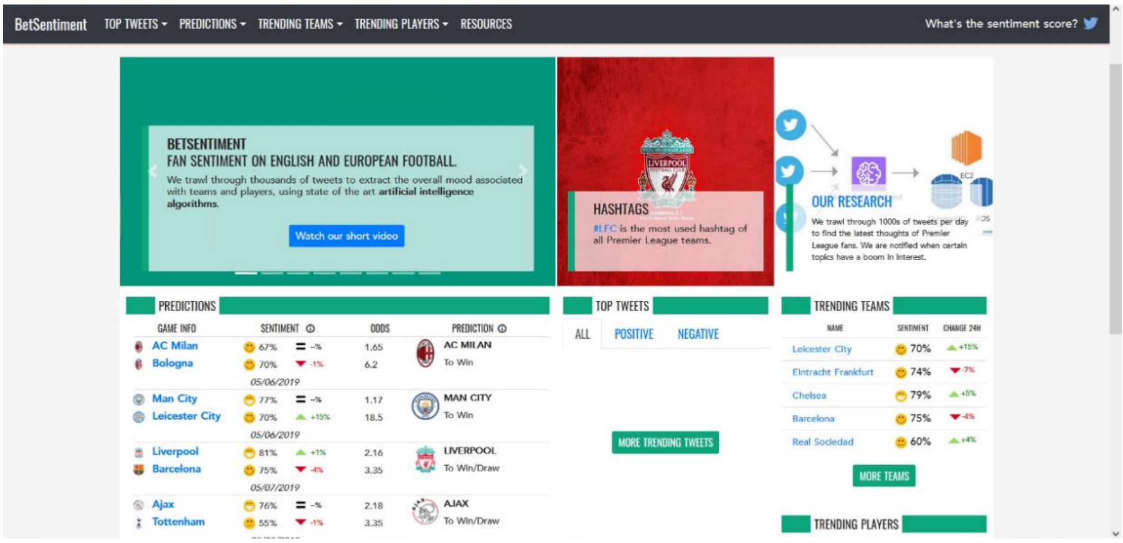
This tool, which requires user registration, works by providing its users with the ability to create monitorization channels, with each channel consisting in a query that filters the collected tweets. These queries can be simple, monitoring tweets referring to a single user, user citation, hashtag, or keyword, or complex, referring to a combination of simple queries using operators (e.g.: and, or, from).

Additionally, through its dashboard, registered users can manage their channels, visualize the analysis and statistics of both their and public channels, and, under request, download collected data sets.

2.3.3 BetSentiment

BetSentiment[25] is a free web-application, currently no longer updated, which focused on providing predictions of football games' winning teams, as well as sentiment analysis on football players and teams, through the ratio of positive and negative tweets from the past 3 days (BetSentiment, 2019).

This web-app, shown in Figure 21, provided an analysis of tweets, based on a ML approach, relative to teams and players, as well as predictions of games from the Champions League, Premier League, La Liga, Ligue 1, Serie A, and Bundesliga.



BETSENTIMENT
FAN SENTIMENT ON ENGLISH AND EUROPEAN FOOTBALL.

We trawl through thousands of tweets to extract the overall mood associated with teams and players, using state of the art **artificial intelligence algorithms**.

HASHTAGS

#LFC is the most used hashtag of all Premier League teams.

OUR RESEARCH

We trawl through 1000s of tweets per day to find the latest thoughts of Premier League fans. We are notified when certain topics have a boom in interest.

PREDICTIONS

GAME INFO	SENTIMENT		ODDS	PREDICTION
AC Milan	67%	= -%	1.65	AC MILAN
Bologna	70%	▼ -1%	6.2	To Win
	05/06/2019			
Man City	77%	= -%	1.17	MAN CITY
Leicester City	70%	▲ +15%	18.5	To Win
	05/06/2019			
Liverpool	81%	▲ +1%	2.16	LIVERPOOL
Barcelona	75%	▼ -4%	3.35	To Win/Draw
	05/07/2019			
Ajax	76%	= -%	2.18	AJAX
Tottenham	55%	▼ -1%	3.35	To Win/Draw

TOP TWEETS

TRENDING TEAMS

NAME	SENTIMENT	CHANGE 24H
Leicester City	70%	▲ +15%
Eintracht Frankfurt	74%	▼ -7%
Chelsea	79%	▲ +5%
Barcelona	75%	▼ -4%
Real Sociedad	60%	▲ +4%

TRENDING PLAYERS

Figure 21 – BetSentiment Website (BetSentiment, 2019)

[25] BetSentiment Website - https://betsentiment.com/

2.4 Discussion of results

The idea of developing a tool for the gathering and analysis of tweets is not a novel concept. With that said, most existing tools resort to the use of lexicon-based methodologies, which are not specialized to the sports' domain, hence, having undesirable performance' levels.

BetSentiment, however, overcomes these limitations by employing a ML-based approach trained on football related tweets, therefore being the tool most similar to the data fusion and analysis framework. Nonetheless, this existing tool significantly differs from the other presented tools and the intended objective by only analysing past tweets, thus deviating from the aim of analysing, in near real-time, sports events' related tweets that were published during the transmission of the respective events. Table 3 exposes a comparison between the presented related works and the proposed framework in more detail.

Table 3 – Comparison between related works and proposed framework

Features	Twitter Sentiment Visualization	Twitter Vigilance	BetSentiment	Proposed Framework
Topic analysis	✓	✓		✓
Lexicon-based sentiment analysis	✓	✓		
ML-based sentiment analysis			✓	✓
Near real-time analysis	✓	✓		✓
Focused on sports' related tweets			✓	✓

With no existing tools capable of fully satisfying the objectives of the proposed framework, the need for the development of a new tool arose. As such, from the study of the SOTA it was possible to identify several promising techniques for, both, the near real-time sentiment and topic analysis.

Regarding the sentiment analysis, the DistilBERT and DistilRoBERTa models show great promise, as they both achieve performance levels comparable to their base models, BERT and RoBERTa, respectively, while being smaller and faster. Furthermore, the DistilRoBERTa model appears to be especially promising as it is derived from the RoBERTa model, which is a model based on the optimization of BERT's pre-training phase. As for the optimization of this models, knowledge embedding techniques prove to be useful, as they are able to improve performance levels. For cases, where labelled data is scarce, the process of reformulating the NLP task into an entailment one, as well as the use of data augmentation also achieve very promising results, being able to compete with supervised techniques trained on much larger datasets.

When it comes to topic analysis, traditional statistical approaches like LDA, NMF, and LSA prove to be effective for keyword discovery, however, when used with the aim of classifying documents based on a set of predefined topics, these techniques fall short when compared to more recent weakly-supervised models, like JoSH or WeSHClass, which allow for the guidance of the learning process. Furthermore, these more modern approaches also improve on the standard statistical models by taking a hierarchical approach to represent the relationships between topics. As for the near real-time aspect, LightLDA achieves very promising results, however like the more traditional approaches it also lacks in the ability to consider a set of predefined topics, hence also suffering from the problem of the learnt topics deviating from the intended topics.

3 Sentiment and Topic Analysis

In this chapter, the dataset used for the training of the models is presented, along with its exploratory data analysis. Additionally, the procedures taken to maintain data security, coupled with the potential risks or ethical violations of the tools to be developed are also detailed. At last, the most promising pre-processing and processing techniques, for, both, the sentiment and topic analysis are presented.

3.1 Exploratory Data Analysis and Description

Analysing the SOTA study results, from both the sentiment and topic analysis, it is possible to conclude that the only common denominator between all presented techniques, either supervised, unsupervised or semi-supervised, is the need for data. As such, for the selection of the pre-processing and processing techniques it is imperative to understand the data which will be used for the task at hand, thus, ensuring the correct selection of techniques.

The following subsection, therefore, presents a description and analysis of the dataset which will be used for the training of the sentiment and topic analysis models.

3.1.1 BetSentiment Dataset

BetSentiment, previously presented in the Text Mining applied to Social Networks section, in addition to the features it offers, also provides a free dataset[26] of close to 8.5 million football

[26] BetSentiment Datasets - https://github.com/charlesmalafosse/open-dataset-for-sentiment-analysis

related tweets, published from May to September of 2018, in the following five different languages: English, Spanish, French, Italian and German.

As the initial aim of the sentiment and topic analysis tools is to analyse English tweets, the exploratory data analysis (EDA) hereby presented focuses only on the dataset's English tweets, which can be divided into, after duplicate removals, 1.920.938 tweets relative to players, 3.462.607 relative to teams and 869.843 relative to the 2018 FIFA World Cup, totalling 6.253.388 tweets. In addition to the tweet's text, each data point of the dataset also contains the tweet's creation date and hour, id, language, sentiment, and sentiment classification score, as presented in Table 4.

Table 4 – BetSentiment dataset's data point

Features	Data Type	Content
tweet_date_created	Continuous	2018-07-07T16:52:01.865000
tweet_id	Discrete	1015639586750500865
tweet_text	Unstructured	@DelMody Brilliant stuff. I was there a couple of months ago &amp; it didn't look like the sort of place I wanted to be. Now, I wish I had been there today!\n\nI raced home from Swansea to Ashford, Kent, to get the game in on the telly. So glad I left early &amp; made it in time.\n\nGo @England!
language	Nominal	en
sentiment	Ordinal	POSITIVE
sentiment_score	Continuous	'{"Neutral":0.00377845601178705692291259765625,"Negative":0.008986658672690391540527343 75,"Positive":0.956345140933990478515625,"Mixed":0.03088975138962268829345703125}'

Despite each data point being composed of six features, only the tweets' text and sentiment are of use for the intended tasks, as such, the EDA of the dataset focused only on these two features.

Regarding the sentiment, the dataset tweets are categorized as either negative, neutral, positive, or mixed, for when a tweet expresses both a positive and negative polarity. This classes, after the removal of the only data point missing the tweet's text, have, respectively, 413.596, 4.413.997, 1.379.604, and 46.190 tweets, thus making this dataset unbalanced, as depicted in Figure 22.

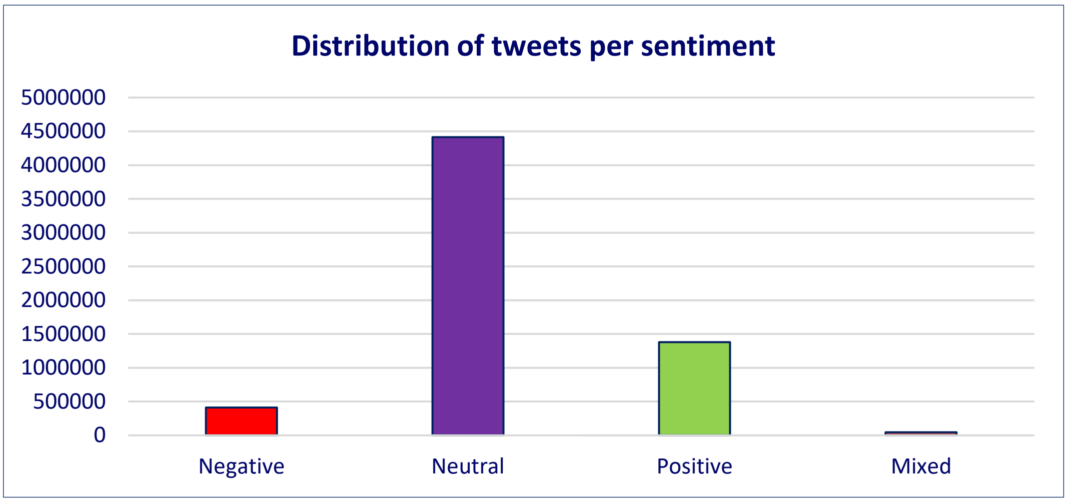


Figure 22 – Distribution of English tweets per sentiment

As a note, it is of importance to mention that the polarity labelling of the tweets was done by the sentiment analysis module of the AWS Comprehend API[27], as such, although able to be used for the training, for the most accurate assessment of the performance of the models, a validation and test dataset must be manually labelled.

As for the tweets' text, an analysis on the word count, unique word count, Uniform Resource Locators (URL) count, hashtag count, mention count, emoji count and most common unigrams was performed, with the results displayed, respectively, in Figure 23, Figure 24, Figure 25, Figure 26, Figure 27, Figure 28, and Figure 29. Analysing the mentioned figures, it is possible to conclude that negative and mixed tweets are, on average, lengthier, in terms of both words and unique words, despite having a smaller presence of URL. When it comes to the presence of hashtags, all classes are distributed fairly equally, however, for mentions it is observable that positive tweets are more inclined to tag only one account, whereas neutral tweets are more inclined to tag two accounts. Negative and mixed tweets, on the other hand, have a more distributed mentions count. Mixed tweets also have a bigger presence of emojis, as opposed to neutral tweets who have the least use of emojis.

[27] AWS Comprehend API Overview - https://docs.aws.amazon.com/comprehend/index.html

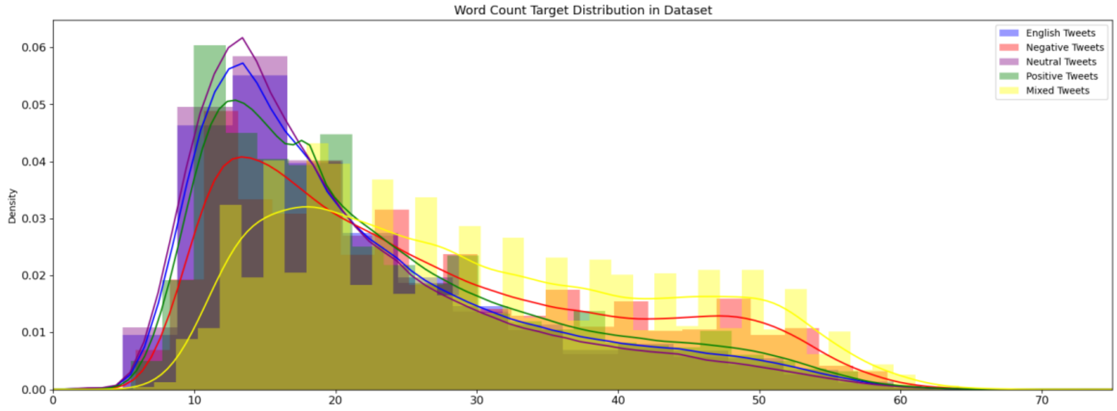


Figure 23 – Word Count Distribution

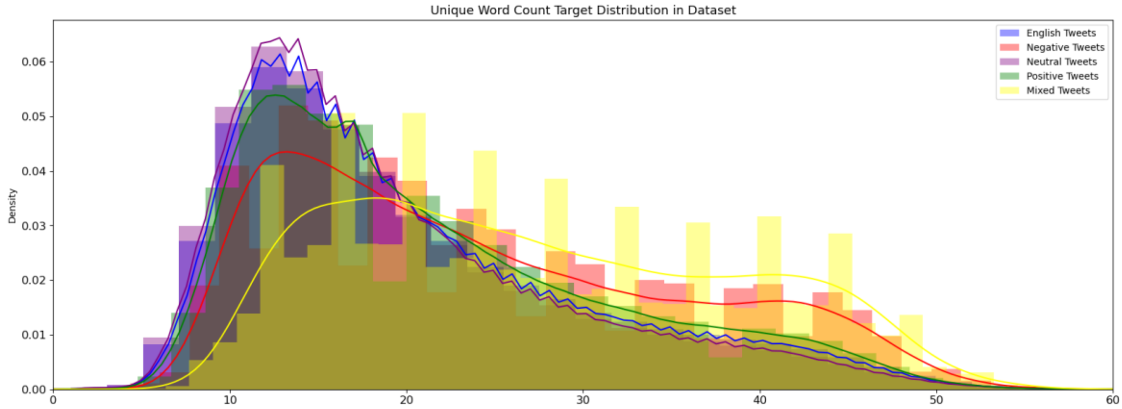


Figure 24 – Unique word count distribution

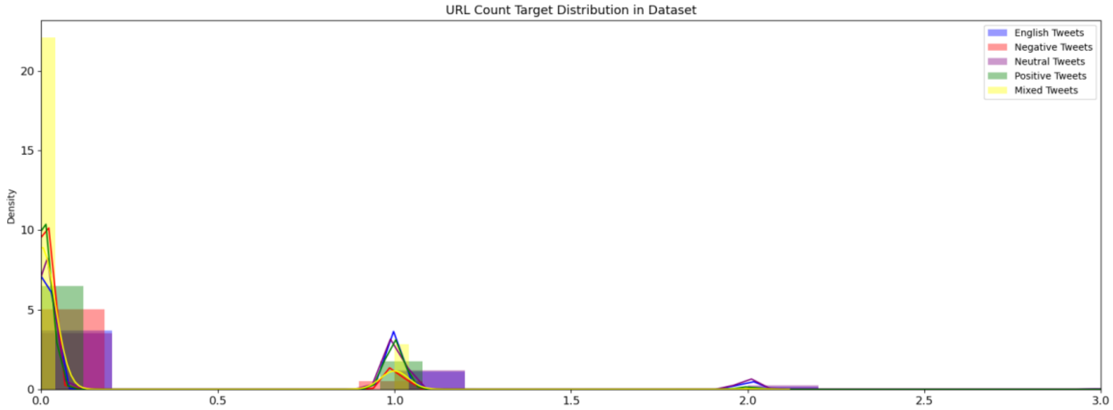


Figure 25 – URL count distribution

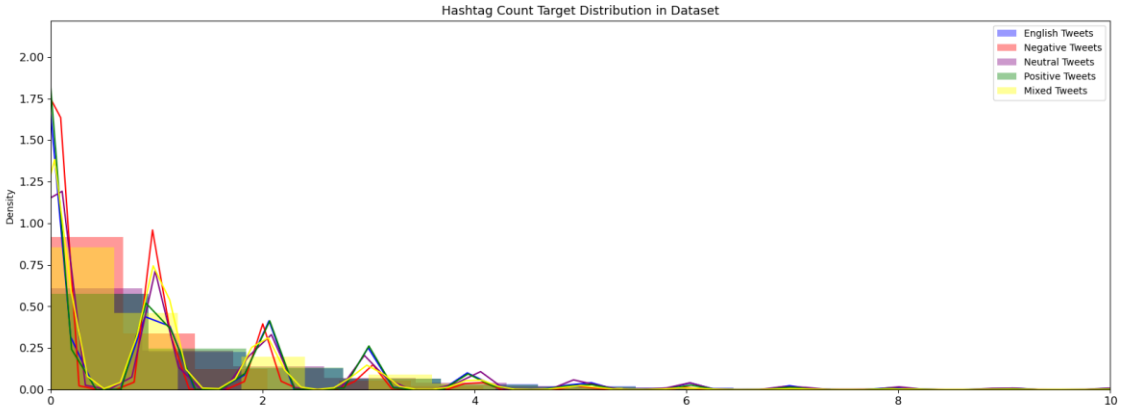


Figure 26 – Hashtag count distribution

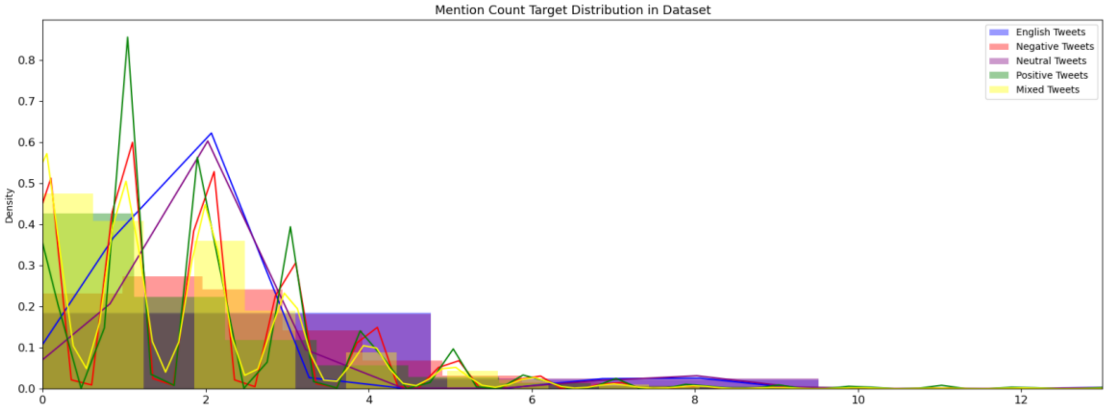


Figure 27 – Mention count distribution

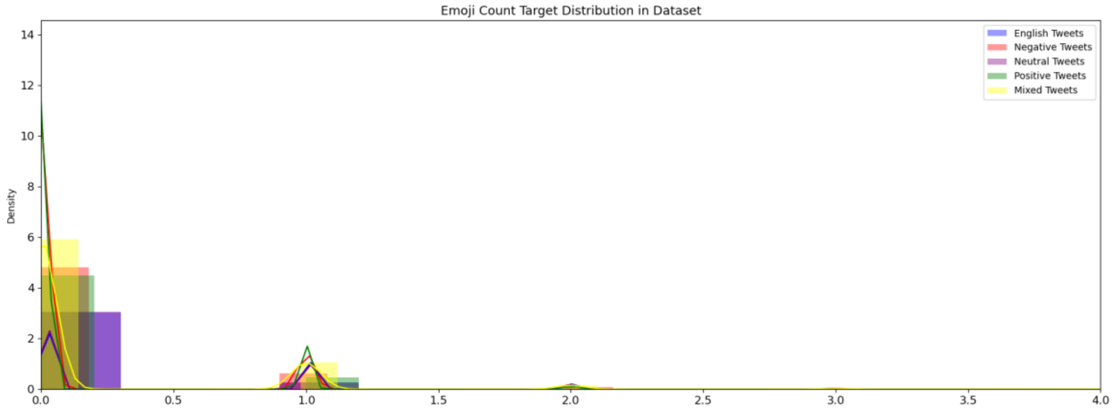


Figure 28 – Emoji count distribution

At last, in regard to the most common unigrams, which are one-word sequences, it is feasible to conclude that not only words can be associated to certain classes, such as “not” with negative tweets and “good” with positive tweets, but also mentions. As such, in Figure 29, it is possible to observe that the Manchester United handle (@manutd) is a common occurrence in the negative tweets.

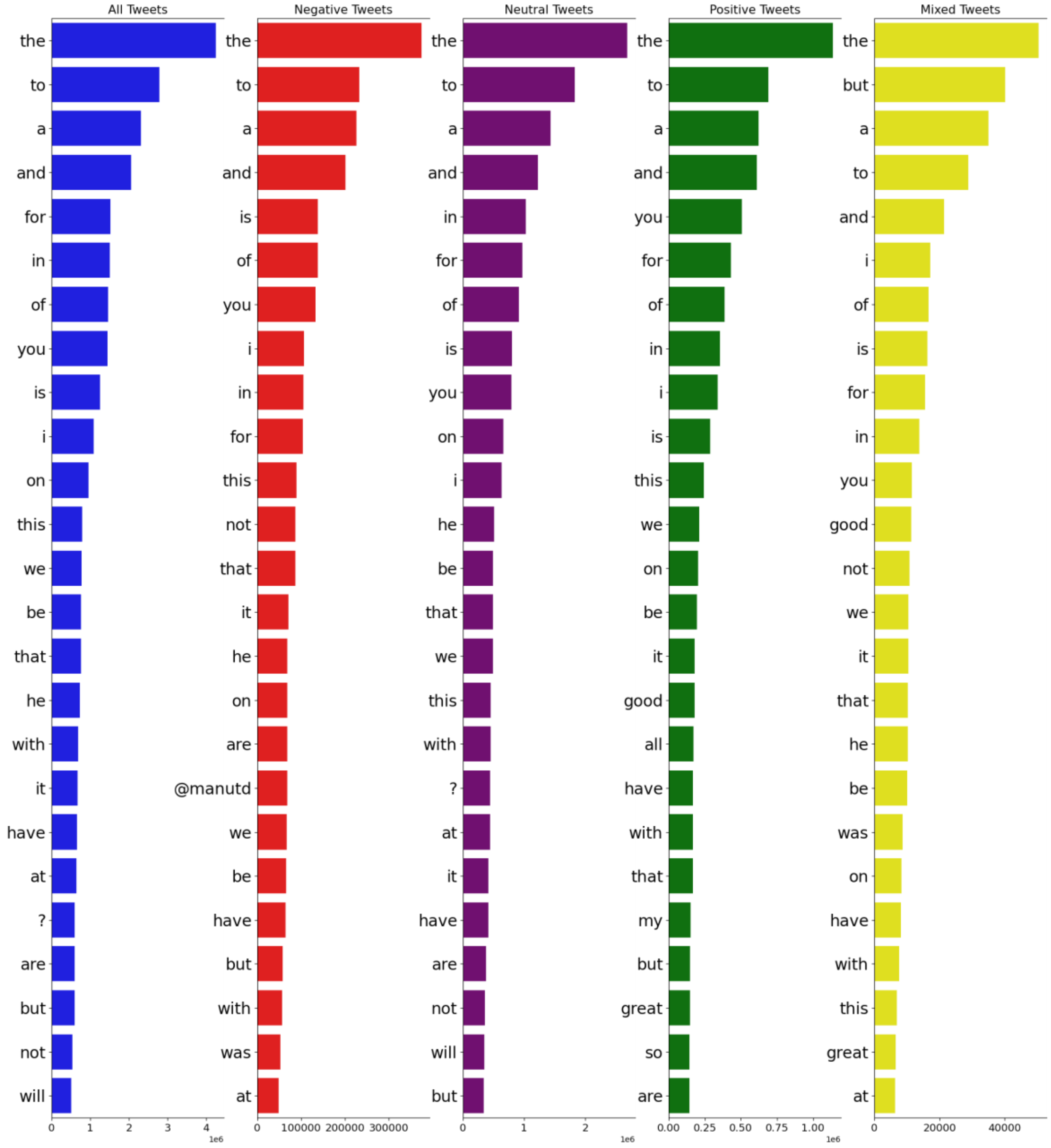


Figure 29 – Top 25 most common unigrams per class

3.2 Data Protection, Security Analysis and Ethical Aspects

As the field of ML evolves and, consequently, its algorithms improve in performance, its application has become more prominent across a wide range of fields, to the point where most people, unknowingly to them, vastly depend on this type of algorithms on a daily basis.

This increasingly dependence, coupled with the powerful abilities of ML models and the often sensitive and confidential information present in the data, however, raises concerns in regard to its safety of use. As such, the following subsections focus on presenting the steps done to ensure the safety of the data, as well as any potential risks or ethical violations of the tools to be developed.

3.2.1 Data Protection

Data protection, more than ensuring the privacy of the data, is the process of safeguarding the data from being corrupted, compromised, or lost (Crocetti et al., 2021), thus guaranteeing its ability to be restored in case of need.

With this in mind, for the development of the project, and considering the limitations of Git[28], a version control system for code files, in handling large files, such as the ones of datasets and machine learning models, it is used DVC[29], which stands for Data Version Control. DVC is a Git-compatible, storage agnostic tool that allows for the version control of large files, allowing changes in the datasets and models to be securely tracked and stored, similarly to how Git tracks changes in code files. In addition, DVC allows for an easy reproducibility of end-to-end experiments, as well as metric tracking, therefore, not only ensuring data protection, but also facilitating the process of fast experimentations.

3.2.2 Security Analysis and Ethics Issues

It is often said that “with great power comes great responsibility”. As such, it is of utmost importance that when developing a tool, all potential risks caused by its incorrect functioning are considered. Due to the nature of ML, it is rare for a model to produce the correct output all the time. As a result, in this field it is especially important that the assessment of all possible risks is done, therefore, helping avoid any negative outcomes in cases where models do not work as expected.

Having said that, in regard to the sentiment and topic analysis models to be developed, as their results only use is to help brands better understand sports events audiences’ perception of the

[28] Git Website - https://git-scm.com/
[29] DVC Website - https://dvc.org/

events, through their sentiments and most discussed topics, there are no potential security issues. As for potential ethics issues, analysing the ICDT Ethics Self-Assessment Guide[30], it is also possible to conclude that the aforementioned tools and the data used for its training do not violate any ethics' rules, as the tweets used do not contain any association to specific accounts, and their gathering and use is in compliance with Twitter Terms of Service.

3.3 Pre-processing

Statistical and ML models' outputs are highly dependent on the quality of the data provided to them. As such, it is of utmost importance that the data used is simplified and formatted according to the needs of the models used.

For the pre-processing of the BetSentiment Dataset, considering its EDA, there were four types of pre-processing techniques used, respectively, for the sampling of the data, its simplification, its splitting, and its formatting. Figure 30 depicts the different pre-processing flows depending on whether the data is labelled, partially labelled or completely unlabelled. As such, for the sentiment analysis, due to the dataset having the sentiment labels for all tweets, all four types of techniques are of use, whereas for the topic analysis, since the dataset has no topic labels, only the data simplification and formatting are of use. Additionally, in case part of the data is manually annotated in regard to the topics, the dataset may also be split for topic analysis evaluation purposes.

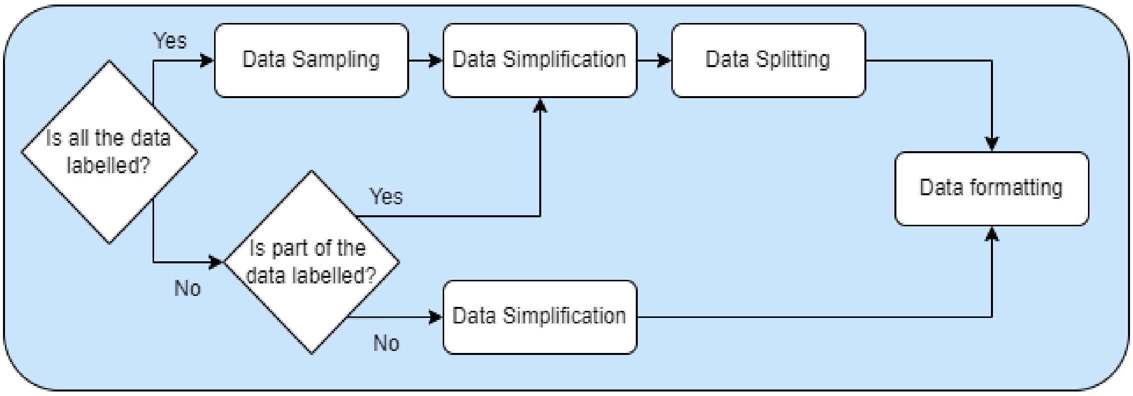


Figure 30 – Pre-processing flow

[30] ICDT Ethics Self-Assessment Guide - https://www.fct.pt/apoios/projectos/concursos/ICDT/docs/ICDT_Ethics_Self_Assessment_Guide.pdf

3.3.1 Data Sampling

From the EDA of the dataset, it was possible to observe that the sentiment classes are unbalanced. As a result, this part of the pre-processing aims to balance the dataset through the use of, either the downsampling of the majority sentiment classes, or the upsampling of the minority sentiment classes. In other words, the tweets are either removed or duplicated to help balance the dataset.

In addition, as the sentiment analysis only aims to classify tweets as positive, neutral, and negative, this pre-processing component is also responsible for the removal of mixed sentiment tweets.

3.3.2 Data Simplification

As implied by the name, this pre-processing component seeks to simplify the data in order to reduce its complexity. As such, considering the EDA of the dataset, 17 simplification steps were implemented, with all except the first step aiming to simplify the tweets' text feature, due to it being the only dataset feature used as input by, both, the sentiment and topic's classification.

Table 5 focus on presenting each of the steps in order of application. Different combinations of this simplification steps are then experimented for both classification tasks with the results presented in the Experimentation of Techniques section.

Table 5 – Pre-processing simplification steps

Simplification steps	What it does
Handling of missing values	Removes data points that have missing values in the selected features. Considering that the only features of the dataset used are the tweets' text and sentiment, this step removes data points that have at least one of this features' value missing.
Handling of HTML characters	Converts all HTML characters, such as, for example, "%20" or "&amp", to its ASCII form.
Removal of handles/mentions	From the EDA it was concluded that mentions can carry a significant, but undesirable connotation. As such, this step removes all handles/mentions from the tweets' text.
Handling of hashtags	Hashtags can, not only consist in words (e.g.: "#winner"), but also phrases, where the words are concatenated (e.g.: "#WeAreTheWinners"). As such, for this step, two approaches were tried. First, removing all hashtags entirely, and secondly, only

Simplification steps	What it does
	removing the hashtag symbol and segmenting the words (e.g.: "#WeAreTheWinners” becomes “We Are The Winners”).
Handling of URLs	From the EDA, it is possible to conclude that there is a slight variation in the presence of URL among tweets from different polarities. As such, for this step it was tried the removal of URL, as well as the replacement of them with a URL token (“[URL]”).
Handling of contractions	Expands contractions (e.g.: “I’ll” becomes “I will”)
Handling of laughs	For this step it was tried the removal of all resembles of laughs, such as, for example, “ahaha” or “lol”, as well as the replacement of them with a laugh token (“[LAUGH]”).
Handling of emojis and emoticons	For this step it was tried, both, the removal of all emojis and emoticons, as well as the replacement of them with tokens which represent their respective emoji or emoticon (e.g.: “ 👍 ^-^” becomes “[thumbs_up] [happy]”).
Removal of repeating characters and punctuation	All punctuation is removed and characters which appear more than two times in a row are reduced to only two characters (e.g.: “Gooaaaaaal.” Becomes “Gooaal”). This reduction to two characters instead of one is done so that these words can be distinguished from the correct words, as they may have a different connotation.
Removal of accents	Removes all accents and converts specials characters to their closest ASCII format (e.g.: “Góoool “Ø” becomes ‘Gooool “O’)
Handling of upper case	All characters are lowercased.
Removal of team names	Team names, such as mentions, can carry undesirable connotations. As such, this step removes football team names from the tweets’ text.
Removal of numbers	All numbers are removed.
Handling of misspelled words	For this step, two approaches to correct misspelled words were tried. The first based on the Jaccard distance (Niwattanakul et al., 2013) and the second based on a Symmetric Delete spelling correction algorithm (Garbe, 2015).
Removal of stop words	All stop words, which are frequent words with little to no significant meaning (K. & R., 2016), are removed. For this step, two stop words’ dictionaries were tried. One was the NLTK[31] stop words dictionary, while the other was a custom dictionary with the most frequent unigrams of the dataset.

[31] NLTK Website - https://www.nltk.org/

Simplification steps	What it does
Removal of words based on part of speech tags	This step removes words based on their part of speech tags (e.g.: noun, verb, determiner, etc.). As such, for this step, using NLTK part of speech tagger[32], it was tried the removal of words whose tags appeared to be of little importance for the tasks at hand.
Word simplification	For this step, two word simplification approaches were tested. Stemming, which is a heuristic process for the retrieval of the root form of a word, also known as stem, (Manning et al., 2009), and lemmatization, which retrieves the base form of a word, also known as lemma, through a vocabulary and morphological analysis (Manning et al., 2009).

3.3.3 Data Splitting

When evaluating a model, it is imperative that the data used for its evaluation has not been used for its training, as such could lead to biased results. As a result, this part of the pre-processing aims to split the data passed to the model into different sets, using either the Holdout method, which splits the data into a training and test set (Schneider, 1997), or an improved Holdout method, which, besides the training and test set, also creates a third set to be used along the training for validation purposes.

3.3.4 Data Formatting

All models have a required input format. As a result, this pre-processing step only function is the one of formatting the data so that it can be used by the intended model, whether through the tokenization, padding, truncation, embeddings, which are vector representations of text, or any other formatting technique of the text data.

3.4 Sentiment Analysis

From the SOTA study, it is possible to conclude that there are multiple promising techniques for the near real-time sentiment analysis task. As such, this subsection focuses on presenting in a greater depth those techniques, along with the changes done to them prior to the experimentation.

[32] NLTK Part Of Speech Tagger - https://www.nltk.org/api/nltk.tag.html

The following presented techniques are the DistilBERT, DistilRoBERTa, and the Knowledge embedding used by such models.

3.4.1 DistilBERT

DistilBERT is a ML model derived from the knowledge distillation of the BERT model (Sanh et al., 2019). As such, in order to fully understand this model, one must first understand the BERT model.

BERT is a language model, pre-trained on the Wikipedia and Google's BooksCorpus, that innovates by being bidirectionally trained, as opposed to the previous approaches which, either train from left to right, right to left, or both. This is done through the use of two innovative training approaches named NSP and masked-language modelling (MLM) which, respectively, train the model through the prediction of if a sentence follows another sentence and the prediction of randomly masked words in a sentence (Devlin et al., 2019). This forces the model to use the words on either side of the masked word to help with the prediction, thus learning in a way more similar to humans.

With this pre-training approach, BERT is then able to be finetuned for a variety of NLP tasks, as represented in Figure 31, with less task-specific data than would be needed otherwise, thus leading to an increase in performance which was able to obtain SOTA results in 11 NLP tasks (Devlin et al., 2019).

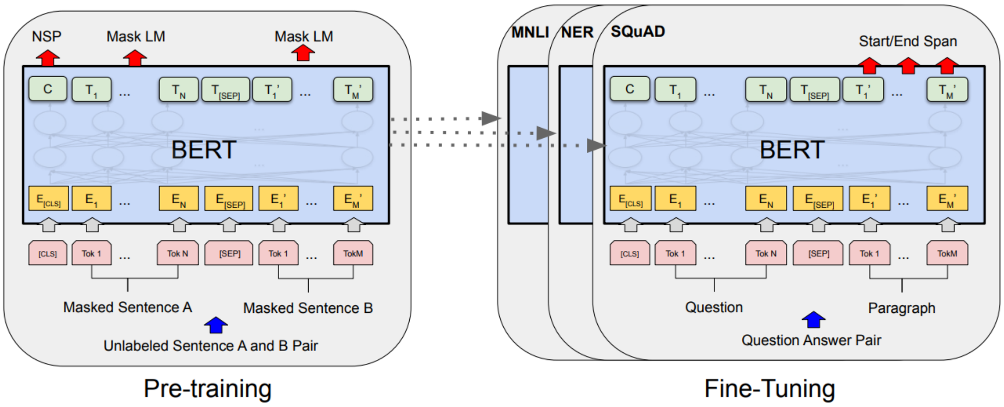


Figure 31 – Pre-training and fine-tuning procedures for BERT. Retrieved from (Devlin et al., 2019)

In addition to the pre-training approaches used, another key factor of BERT's success is its architecture, which as depicted in Figure 32, is composed of 12 stacked Transformer encoders

(Devlin et al., 2019). Thus, taking advantage of their parallelizable architecture, as well their attention mechanisms, which allow each encoder to focus on the more relevant features of their input data (Vaswani et al., 2017).

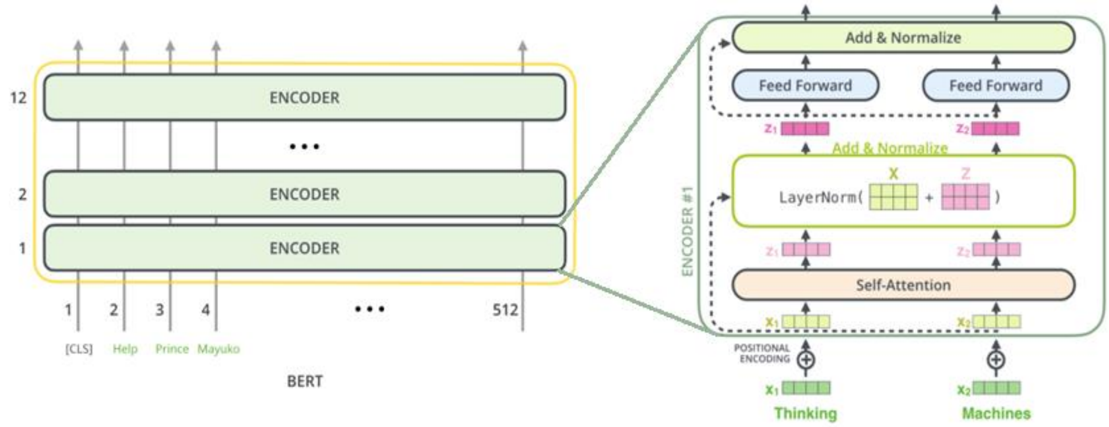


Figure 32 – BERT architecture. Adapted from (Alammar, 2021)

This, despite leading to faster training and inference times than more conventional approaches which rely on convolutions and recurrences, still falls short in the domain of near real-time use. Reason why, aiming to improve on this aspect, DistilBERT was introduced (Sanh et al., 2019).

By improving on BERT through the removal of token-type embeddings and the pooler, along with the reduction in the number of layers in half, this new model is able to shrink by 40%. Additionally, through the optimization of its linear algebra equations using modern frameworks, along with its initialization using the weights of one of each two layers of BERT, it is able to achieve 60% faster times, while preserving 97% of the language understanding capabilities of its predecessor (Sanh et al., 2019).

For the fine-tuning of this model for the sentiment analysis task, as presented in Figure 33, two additional layers were added to DistilBERT, a dropout layer, which by randomly dropping out nodes during training aims to make the model more robust, and a fully connected linear layer. Regarding the output format, considering the lack of need for the probabilities of each label, it was decided for the use of the argmax function on the output, rather than the use of the softmax, due to it being faster (Li, 2019).

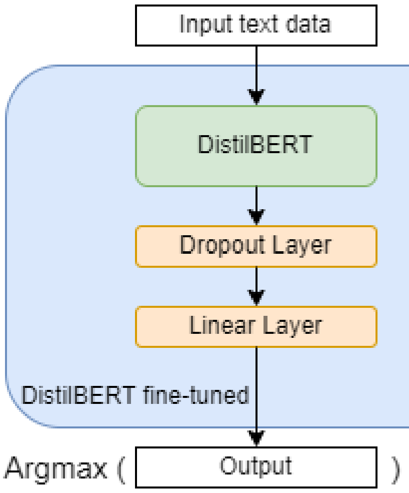


Figure 33 - DistilBERT fine-tuning architecture

At last, in addition to the added layers and the output function, for the fine-tuning of the model, the following hyperparameters were also defined:

- **Loss function = Cross-entropy loss –** The loss function is responsible for quantifying the difference between the expected result and the actual result. Due to the model being used for a multi-classification problem, the loss function used was the cross-entropy loss.
- **Optimizer = Adam –** The optimizer is responsible for minimizing the loss function. The optimizer used was Adam (Kingma & Ba, 2015), due to its faster running time, low memory requirements, less tuning requirements, and ability to update the learning rate (A. Gupta, 2021).
- **Learning rate = 3e^{-5} –** The learning rate is responsible for defining the magnitude of changes to weights during the backpropagation training process (Yi Li, 2021). The learning rate used was 3e^{-5}, due to it being one of the learning rates used in BERT's original paper (Devlin et al., 2019).

3.4.2 DistilRoBERTa

DistilRoBERTa follows the same logic of DistilBERT, however with RoBERTa as the base model (Hugging Face, 2019). As a result, given that the pre-training approach is the sole difference between BERT and RoBERTa, the only aspect of this model that distinguishes it from the model that was previously presented is the pre-training weights. As such, in order to fully understand this model, one must first further understand the difference between BERT and RoBERTa.

RoBERTa improves on BERT by modifying its pre-training approach through the removal of the NSP task and the changing of the MLM task to use dynamic masking patterns. This, combined with the pre-training on more data, utilizing bigger batches and longer sequences, led the model to outperform its predecessor, thus obtaining SOTA results on GLUE, RACE, and SQuaD (Y. Liu et al., 2019).

Regarding the fine-tuning process, due to the similarity with the DistilBERT model, it is applied the same modifications described in the previous subsection.

3.4.3 Knowledge embedding

Knowledge embedding is the process of adding previously known knowledge to a model. As such, for the knowledge embedding of both presented sentiment analysis' models, following the technique presented in (Ostendorff et al., 2019), it is used knowledge graph embeddings of Wikipedia pages.

As a result, prior to the use of this technique, and having been decided for the enriching of the models with football teams' Wikipedia pages, it was used a scraper for the gathering of the pages, allowing, consequently, for the retrieval of their embeddings from the Facebook's PyTorch-BigGraph model (Lerer et al., 2019) trained on the Wikidata[33] graph. Having the embeddings, the dataset tweets were then associated with teams by comparing the tweets' date and time of creation with football game schedules, scraped from Flashscore[34], of the English Premier League, Champions League and 2018 World Cup, as well as by comparing the games' teams with the teams' mentions or hashtags in the tweets. This led to a total of 819.168 tweets, which were published during the 1891 games scraped, to be annotated with their respective game and team.

Having, both, the embeddings and the tweets associated with teams, for the fine-tuning of the models using this technique, as depicted in Figure 34, the non-fine-tuned output of the model is concatenated with the embedding associated to that tweet's team. However, because not all tweets may have a team, or not all teams may have a Wikipedia page, and thus embeddings, a Boolean representing if a real embedding is being used or not is also passed to the

[33] Wikidata Website - https://www.wikidata.org/wiki/Wikidata:Main_Page

[34] Flashscore Website - https://www.flashscore.com/football/

concatenation. The concatenation result is then processed through two fully connected layers with 1024 neurons, ending in an output layer, on which the argmax function is applied.

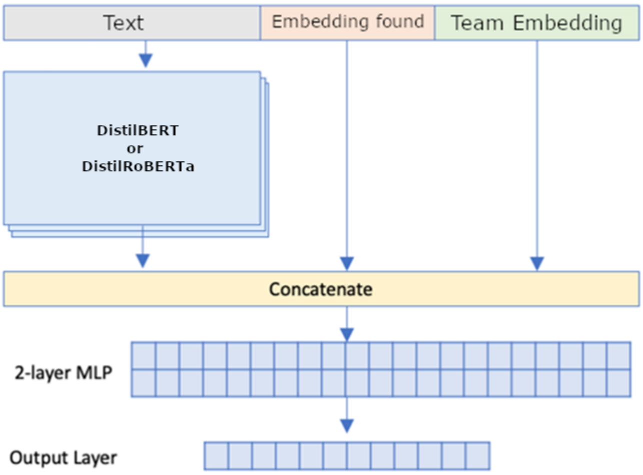


Figure 34 - Knowledge embedding architecture. Adapted from (Ostendorff et al., 2019)

3.5 Topic Analysis

From the SOTA study, it is possible to conclude that there are multiple promising techniques for the near real-time topic analysis task. As such, this subsection focuses on presenting in a greater depth those techniques.

The following presented techniques are the JoSH and WeSHClass models.

3.5.1 JoSH

JoSH as represented in Figure 35, is a topic mining model introduced with the aim of discovering hierarchical structured topics in a guided manner, through the use of a category tree composed only of the topic words (Meng et al., 2020).

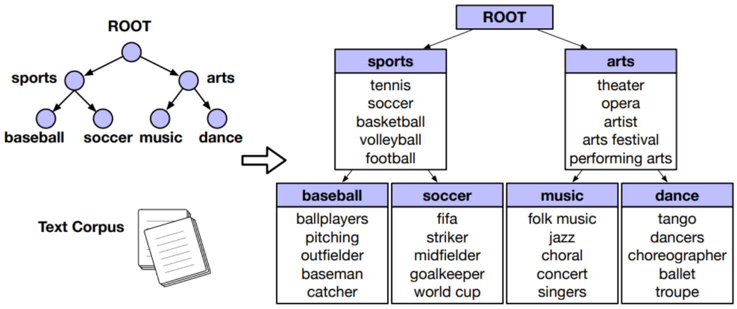


Figure 35 – Hierarchical topic mining. Retrieved from (Meng et al., 2020)

As such, this model aims to discover the most relevant terms for each topic, while preserving their relationships in the spherical embedding space, by considering, both, the intra-category and inter-category coherence (Meng et al., 2020). This, as represented in Figure 36, allows the model to represent the terms of a topic close in the embedding space, while still maintaining distinctiveness between topics and while preserving the relative distance within local trees, meaning that the tree distance between two children nodes is larger than that between a children node and the parent node (Meng et al., 2020).

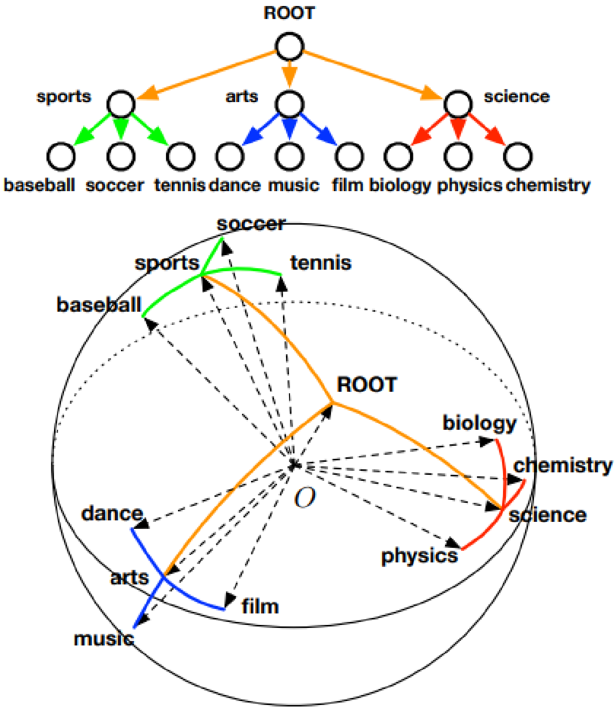


Figure 36 – Spherical tree embeddings. Retrieved from (Meng et al., 2020)

In more detail, this model functions by matching each text corpus document with one of the topics, then modelling the semantic coherence between a word and the document in which it appears, and, at last, modelling the semantic correlation of words that co-occur inside a local context window (Meng et al., 2020). Additionally, due to the embedding space used being spherical, rather than the more traditional Euclidean optimization techniques, it is employed the Riemannian optimization to optimize the learning of the embeddings (Meng et al., 2020).

As for the topic classification of documents, due to the explicit assumption that topics and documents are generated from one another, this model makes it possible to build a generative classifier that places the document in the category where there is the greatest chance that it will be generated from (Meng et al., 2020).

3.5.2 WeSHClass

WeSHClass is a neural model for hierarchical text classification, which uses weak supervision to help reduce the amount of data necessary for the training. As such, by modelling each class, using, either a set of keywords for each topic or topic-classified documents, coupled with the use of a Bag-of-Words (BOW) or LSTM model trained on the dataset, it is able to generate pseudo documents to be used for the pre-training of the model (Meng et al., 2019).

In more detail, the pseudo document generation works by modelling each class using a mixture of von Mises Fisher distributions, which are spherical probability distributions, (Mardia, 1975) fitted with, either the closest words in the embedding space to the average embedding of the class keywords or, when provided with topic-classified documents, the words of the labelled documents with the highest term frequency-inverse document frequency (TF-IDF) (Meng et al., 2019). This allows for the selection of a topic-related word to be used as the beginning of each pseudo document, thus enabling the rest of the document to be generated by a BOW or LSTM model trained on the dataset (Meng et al., 2019).

As for the embeddings, even though the model uses a high-dimensional spherical embeddings space, because of its capability to normalize the vectors so that they reside on a unit sphere, it has support for more conventional word embedding techniques, such as Word2Vec (Mikolov et al., 2013), which, as presented in Figure 37, provides two methods for the learning of embeddings. Continuous Bag-of-Words (CBOW), which attempts to predict a word based on its context, therefore being faster and having better representations for frequent words, and Skip-gram, which given a word tries to predict its context, thus working better with a smaller amount of data and being better at representing infrequent words (Dhruvil Karani, 2018).

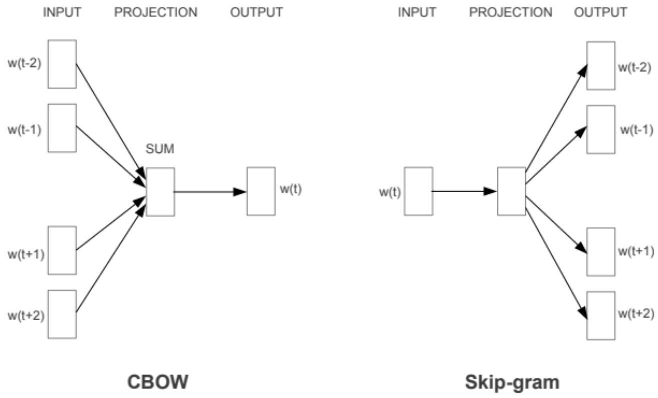


Figure 37 - CBOW and Skip-gram model architectures. Retrieved from (Mikolov et al., 2013)

Additionally, this model also allows for the topic hierarchy to be provided, thus enabling the use of a hierarchical structure of deep neural networks, depicted in Figure 38, that due to mimicking the given hierarchy and having a blocking mechanism, is capable of selecting the appropriate levels for documents (Meng et al., 2019).

For the ability of the model to perform the hierarchical topic classification, however, it is necessary that each hierarchy has the probability distribution spread over all topics. As such, besides the local classifiers, a global classifier is trained on the unlabelled data for the assignment of document soft probabilities at each level, with the final topic prediction, as presented in Figure 38, being made through the multiplication of all classifiers' outputs from the root to the current levels, thus giving lower-level classifiers chances to correct misclassifications made at higher levels (Meng et al., 2019).

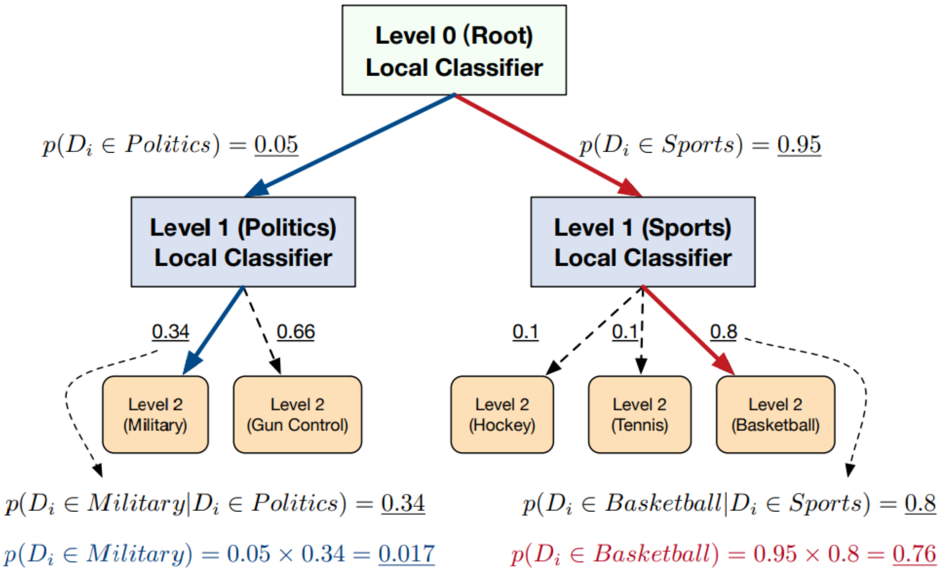

Level 0 (Root)
Local Classifier
$p(D_i \in Politics) = \underline{0.05}$
$p(D_i \in Sports) = \underline{0.95}$
Level 1 (Politics)
Local Classifier
Level 1 (Sports)
Local Classifier
0.34
0.66
0.1
0.1
0.8
Level 2
(Military)
Level 2
(Gun Control)
Level 2
(Hockey)
Level 2
(Tennis)
Level 2
(Basketball)
$p(D_i \in Military | D_i \in Politics) = \underline{0.34}$
$p(D_i \in Basketball | D_i \in Sports) = \underline{0.8}$
$p(D_i \in Military) = 0.05 \times 0.34 = \underline{0.017}$
$p(D_i \in Basketball) = 0.95 \times 0.8 = \underline{0.76}$

Figure 38 – Hierarchical neural structure. Retrieved from (Meng et al., 2019)

3.6 Discussion of results

From the SOTA study, it is possible to conclude that, regardless of the task, the dataset must be chosen prior to the selection of the models. As a result, after a thorough search of datasets for the sentiment and topic analysis of football related tweets, the BetSentiment Dataset proved to be the best, due to having the sentiment labels for all tweets, while also having the biggest and more diverse set of tweets of the datasets found, consisting in 6.3 million tweets about teams, players and games of the Premier League, Champions League, and 2018 World Cup.

Having the dataset, however, is not enough for the training of the models, as the data must first be adjusted according to the task at hand and the models' characteristics. As such, along with the selection of the models to be evaluated, four types of pre-processing techniques (data sampling, simplification, splitting, and formatting) were implemented, with their use depending on whether or not the data is labelled.

As for the sentiment analysis models' selection, considering the most promising sentiment analysis SOTA techniques, as well as the planned future support for more data sources and sports, it was decided for the use of the DistilBERT and DistilRoBERTa pre-trained language models, which already have some understanding of the language, thus requiring less data for the training, and consequently allowing for the future use of smaller datasets of other sports and data sources. Additionally, with all the dataset tweets having a mention or hashtag of a team, as well a date and time of creation it was possible to associated them with specific games, as well as with one of the teams in the game, thus allowing for the knowledge embedding of the pre-trained models with the embeddings of the teams' Wikipedia pages.

Regarding the topic analysis, accounting for the lack of topic labels in the dataset and the planned future support for more sports, it was decided for the use of the JoSH and WeSHClass hierarchical topic models, thanks to their support for weak supervision signals, as well as their ability to account for the relationships between topics, thus allowing for a better distinction of topics of different categories.

At last, considering the importance of having the data and the models safely stored, despite the lack of security and ethical concerns, it was decided for the employment of a Git-compatible, storage-agnostic tool, named DVC, which, in addition, to the tracking and storing of both the dataset and the trained models, allows for an easy reproducibility of end-to-end experiments, as well as metric tracking, therefore providing benefits beyond data security.

4 Experimentation of Techniques

In this chapter, the evaluation metrics considered for the measurement of the performance of the models are presented, along with the experiments done to compare the different pre-processing and processing techniques for, both, the sentiment and topic analysis. Additionally, the final trained models for both tasks are also presented.

4.1 Evaluation Metrics

One of the last key steps in the data science process consists in the selection of the model, however, for this it is necessary for the model's performance to be represented in a comparable way. Evaluation metrics, therefore, aim to solve this problem by providing an easy to compare summarization of the quality of statistical and ML models (DeepAI, 2020).

Nonetheless, as with the pre-processing and processing techniques, when choosing the evaluation metric to be used for the representation of the models, one must consider, both, the data used and the model's task. As such, this section focuses on presenting evaluation metrics commonly used with, both, sentiment analysis models and topic analysis models.

4.1.1 Sentiment Analysis

4.1.1.1 Confusion Matrix

A confusion matrix, also known as an error matrix (Stehman, 1997), is a performance measurement for supervised classification problems, consisting in a matrix where the number of rows and columns is equal to the number of classes that the model can predict. In this matrix and following Figure 39, each column represents instances predicted as that column's class, whereas rows represent instances that are actually from that class.

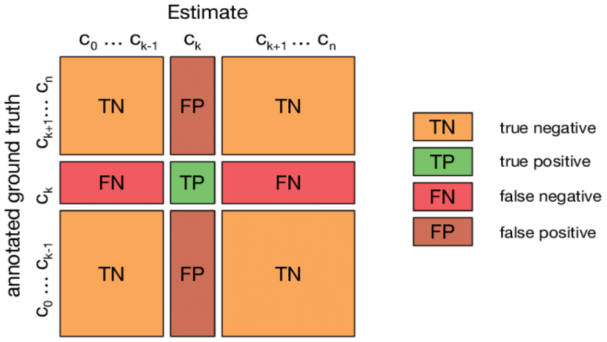


Figure 39 – Multi-class confusion matrix. Retrieved from (Krüger, 2018)

In a confusion matrix, predictions can be categorized as, either, true negative, true positive, false negative, or false positive, thus this type of matrices allow for the exposure of the following values (Google Developers, 2020c):

- **TN**: Number of instances from the negative class(es) correctly classified, known as true negatives.
- **TP:** Number of instances from the positive class correctly classified, known as true positives.
- **FN:** Number of instances incorrectly classified as from one of the negative classes, known as false negatives.
- **FP:** Number of instances incorrectly classified as from the positive class, known as false positives.

4.1.1.2 Accuracy

Accuracy is a metric used to evaluate the proportion of correct predictions compared to the total number of predictions (Dalianis, 2018). As such, this metric results in a value between 0 and 1, where 0 means that no prediction was correct and 1 that all predictions were correct. (1) exposes the formula for the accuracy.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

The quality of this performance measure is strongly associated to the distribution of the classes used for the assessment. As such, in cases where one has an unbalanced dataset, this evaluation metric can be misleading (Google Developers, 2020a).

4.1.1.3 Precision and Recall

Precision and recall are two evaluation metrics that complement each other, therefore, usually, used together. Precision aims to answer “What proportion of positive identifications was actually correct?” (Google Developers, 2020b), thus being used to calculate the ratio of false positives. The formula for the precision is exposed in (2).

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall, on the other hand, aims to answer “What proportion of actual positives was identified correctly?” (Google Developers, 2020b), therefore being used to calculate the ratio of correctly classified positive instances. The formula for the recall is exposed in (3).

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

Both metrics present results in the range of 0 to 1, with high values representing, in case of the precision, a low number of false positives, and in the case of the recall, a high number of correctly classified positive instances among all positive instances.

4.1.1.4 F1 Score

F1 score, also known as balanced F-score or F-measure, is the harmonic mean of the precision and recall metrics (Sasaki, 2007), resulting in 0 as the minimum value and 1 as the maximum. (4) exposes the formula for the F1 score.

$$F1\,Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{4}$$

This evaluation metric, due to being based on the precision and recall, focuses on the false negatives and false positives, thus allowing its use with unbalanced datasets (Huilgol, 2019).

4.1.2 Topic Analysis

4.1.2.1 Perplexity

Perplexity, often used for the measurement of the quality of language models (Campos et al., 2018), but also able to be used for the evaluation of topic models, is a metric based on the model’s likelihood, which represents how surprised a model is with a new corpus. In other words, it is a probability that estimates how likely a new test corpus (W_{test}) is, given the n-gram probabilities of a train corpus. As such, higher likelihood values are better.

(5) presents the formula for the calculation of the likelihood of a n-gram model, where N is the length of the test corpus.

$$Likelihood \ = p(W_{test}) = \prod_{i=1}^{N+1} p(w_i | w_{i-n+1}^{i-1}) \tag{5}$$

Regarding the perplexity, as shown in (6), with the likelihood being in the denominator, the lower the value is, the better the model can be considered.

$$Perplexity = \ p(W_{test})^{-\frac{1}{N}} = \frac{1}{\sqrt[N]{p(W_{test})}} \tag{6}$$

4.1.2.2 Coherence Score

Coherence score is a metric used to evaluate the coherence of the extracted topics from a text corpus. As such, this score results from the aggregation of the topics coherence of a topic model (Kumar, 2018), being a topic coherence a measurement of the degree of semantic similarity between the topic's high scoring words (Kapadia, 2019).

There are, however, several ways of calculating semantic similarity between words, thus, when using this metric, it is necessary for the selection of the coherence measure to use (Kapadia, 2019). The existing coherence measures are:

- **C_v:** Based on a sliding window, one-set segmentation of the high scoring words and an indirect confirmation measure, which uses normalized PMI (Bouma, 2009) and the cosine similarity (Alake, 2020).
- **C_p:** Based on a sliding window, one-preceding segmentation of the high scoring words and the confirmation measurement of Fitelson's coherence (Fitelson, 2003).
- **C_uci:** Based on a sliding window and the PMI of all word pairs of the high scoring words.
- **C_umass:** Based on document co-occurrences counts, a one-preceding segmentation and a logarithmic conditional probability as confirmation measurement.
- **C_npmi:** Equivalent to the C_uci, but using normalized PMI, instead of PMI.
- **C_a:** Based on a context window, a pairwise comparison of the high scoring words and an indirect confirmation measure, which uses normalized PMI and the cosine similarity.

4.2 Sentiment Analysis

In this subsection lie the experiments done in regard to the sentiment analysis tool. As such, this subsection starts by displaying the results of the experiments of the pre-processing's simplification steps, followed by the results of the training of the different models, presented in the Sentiment Analysis section of the previous chapter, with the best combination of steps tested.

As for the evaluation of the experiments, considering the importance of having, both, fast and accurate processing of the tweets' text it was used two metrics. The accuracy, which was coupled with the balancing of the dataset, given the intrinsic complexity of unstructured text, the size of the dataset, and the fact that all classes have the same importance, and the time needed to pre-process or process the data. As such, considering the number of experiments planned and the time it takes to run each one, it was used a balanced sample of the dataset with 100.000 tweets, of which 70.000 are for training, 15.000 for validation, and the remaining 15.000 for testing.

All experiments hereby presented were executed using Python 3.8.13, PyTorch 1.10.0, Catalyst 22.04 and Cuda 11.3.1 on a machine with a NVIDIA GeForce RTX 3080 10GB graphics card, an Intel Core i7-12700K processor and 32GB of DDR4 3600 MHz RAM.

4.2.1 Pre-processing experiments

As the SOTA techniques presented in the preceding chapter's Sentiment Analysis section are primarily based on optimizing BERT, either to make it faster or improve its performance, it was decided that the baseline model should be the one most comparable to it. As a result, DistilBERT was chosen, thus making it the model used for evaluating all the pre-processing approaches.

In total, 16 experiments were carried out, with each either adding a new pre-processing step to a previous experiment or replacing an already tested step. As such, a first experiment, in which all other experiments were based on, consisting in 9 pre-processing steps, was devised, taking into account the EDA of the dataset and the pre-processing used in some of the papers already mentioned in the Text Mining chapter. This base experiment steps, ordered by application, consisted in the conversion of HTML characters into their respective ASCII form, the removal of handles, hashtags, URLs and resembles of laughs, the replacement of emojis and emoticons with their respective tokens, the removal of accents and the lowercasing of all text, as the DistilBERT model used was pretrained on lowercased text.

Figure 40 shows 15 of the pre-processing experiments, including the base experiment, with the only one missing being the experiment for fixing misspelled words using Jaccard distance, which due to having an average pre-processing time per 100 instances of 73.27 seconds was immediately excluded for being too slow for near real-time use.

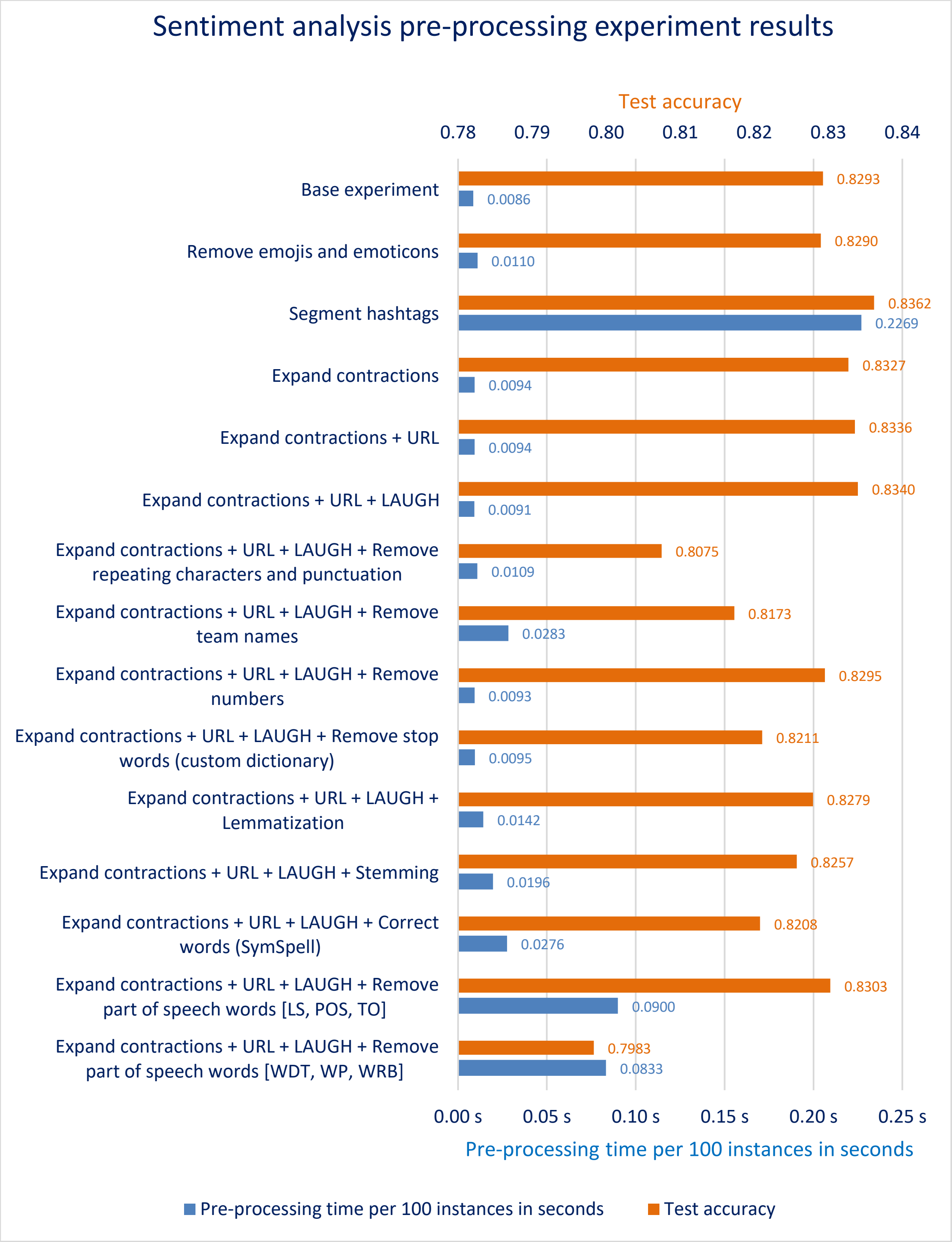


Figure 40 – Sentiment analysis pre-processing experiment results

Analysing the experiments' results, beforehand the selection of the best pre-processing approach, the inference time per 100 instances of the DistilBERT model was acquired, averaging at 0.1663 seconds across all the experiments. As such, taking both metrics into account, the best pre-processing approach was determined to be the one with the best accuracy but a pre-processing time per 100 instances that was inferior to DistilBERT's inference time per 100 instances. As a result, the experiment that built on the base one by expanding contractions, and substituting URLs and resembles of laughs with tokens was determined to be the best pre-processing strategy for the sentiment analysis task.

As a note, it is of importance to mention that due to the data formatting part of the pre-processing being dependent on the model, rather than the data itself, its pre-processing time was included in the model inference time instead of in the pre-processing time.

4.2.2 Processing experiments

With the pre-processing chosen, the models were then tested with the combination of steps that produced the best result, with the only variation in the pre-processing being the keeping of uppercase text for the training of the DistilRoBERTa, due to it being case-sensitive.

ML models, however, have a certain randomness, as such, each model was trained three times, so that the results' average, which are displayed in Figure 41, could be used as the comparison metric, as opposed to the results of single tests, which have a higher likelihood of being skewed.

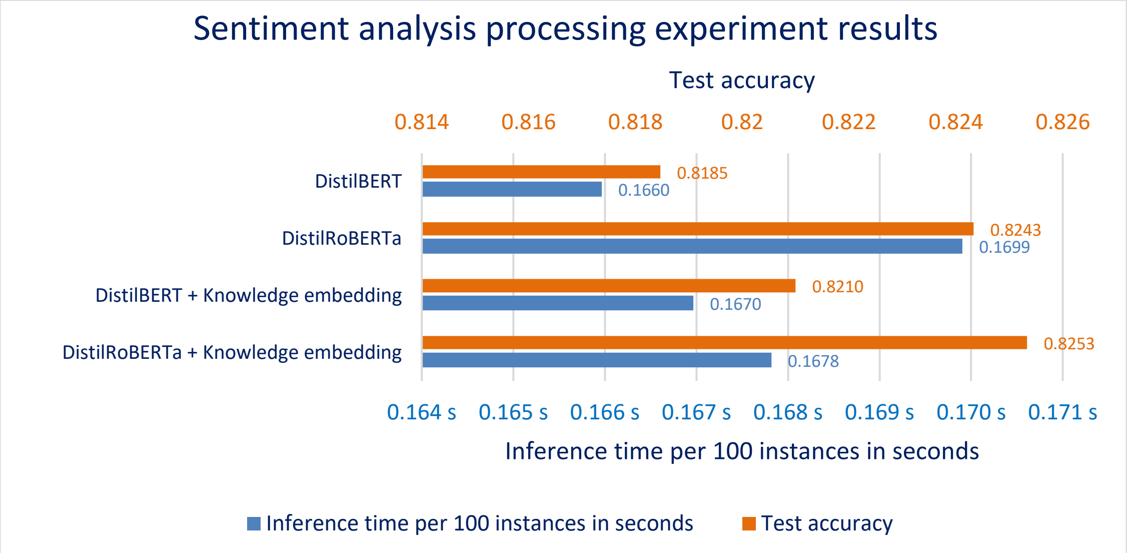


Figure 41 – Sentiment analysis processing experiment results

Analysing the models' results, it is possible to conclude that all inference times per 100 instances are very close, as the difference between the fastest and slowest time is only of 0.0039 seconds. As such, the selection of the model to be trained on the entire dataset was based on the highest accuracy, thus making DistilRoBERTa with knowledge embedding the best tested model for the sentiment analysis task.

4.2.3 Final model

After being chosen, the DistilRoBERTa model with knowledge embedding was then trained on the complete dataset. However, due to the dataset being unbalanced, it was upsampled prior to training, resulting in a total of 13.241.991 tweets, of which 70% were used for training, 15% for validation, and the remaining 15% for testing. The training was subsequently done over 6 epochs, the number of times the dataset was passed to the model for training, with the smoothed accuracy over the training displayed in Figure 42.

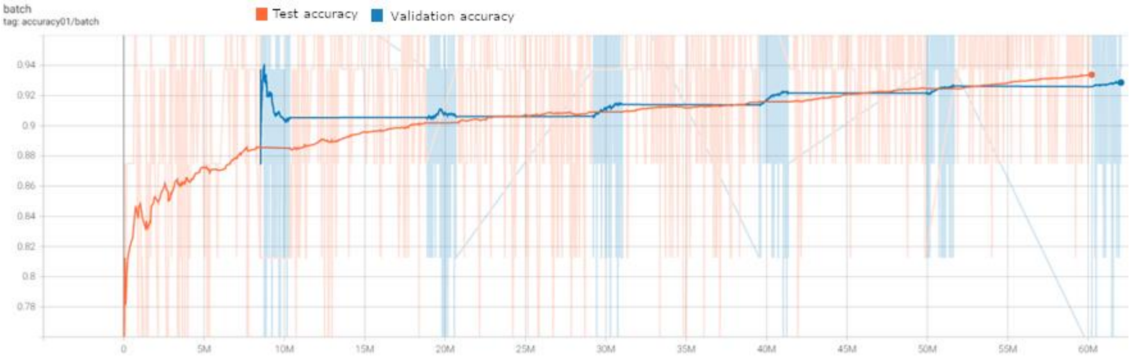


Figure 42 – Sentiment analysis model accuracy

Analysing the accuracy graph, it can be seen that both the validation and testing accuracy were still improving at the sixth epoch, implying that the model was undertrained. As a result, the training was resumed for three more epochs, thus totalling at nine epochs with an accuracy of 94.7%.

4.3 Topic Analysis

In this subsection lie the experiments done in regard to the topic analysis tool. As such, this subsection begins by presenting an initial unsuccessful approach, followed by the results of the pre-processing and processing experiments that led to the selection of the final techniques.

All experiments hereby presented were executed using Python 3.8.13, Tensorflow 2.9.0 and Cuda 11.2.2 on a machine with a NVIDIA GeForce RTX 3080 10GB graphics card, an Intel Core i7-12700K processor and 32GB of DDR4 3600 MHz RAM.

4.3.1 Initial unsuccessful approach

In a first moment, considering, both, the lack of topic labels in the dataset and the topic analysis' objective, which is the one of gathering the main social media' discussed topics of sport events, it was decided for the experimentation of some of the more traditional unsupervised statistical techniques, such as LDA, NMF, and LSA, along with the coherence score as the evaluation metric, in the hope that without any supervision these models would be able to discover useful topics.

This, however, proved to be unsuccessful in two ways. Firstly, these models, due to being focused on topics' discovery, rather than topics' classification, when trained, were only able to provide the keywords of each topic, instead of the topic word itself. This, combined with the lack of supervision led to the creation of topics whose keywords were of hard interpretation, as shown in Table 6, thus making their use impossible.

Table 6 – LDA topic keywords

LDA topic keywords
0.097*"game" + 0.085*"play" + 0.046*"goal" + 0.027*"score" + 0.022*"everi" + 0.020*"half" + 0.020*"ball" + 0.018*"chanc" + 0.015*"defend" + 0.012*"midfield"
0.084*"season" + 0.059*"back" + 0.039*"start" + 0.036*"leagu" + 0.023*"top" + 0.021*"point" + 0.020*"anoth" + 0.020*"transfer" + 0.018*"end" + 0.017*"year"
0.063*"good" + 0.034*"realli" + 0.033*"hope" + 0.025*"alway" + 0.024*"make" + 0.021*"feel" + 0.017*"bad" + 0.014*"sad" + 0.013*"stay" + 0.011*"miss"
0.059*"time" + 0.042*"watch" + 0.029*"year" + 0.020*"week" + 0.015*"work" + 0.015*"thing" + 0.014*"live" + 0.013*"night" + 0.013*"tri" + 0.013*"put"

Secondly, the evaluation metric used, proved to be unsuitable for the task, as it only assessed the quality of the discovered topics, and not the inference quality. As such, in addition to the need of experimenting with new models, there was also the need to switch the evaluation metric, reason why it was decided for the partial annotation of the dataset regarding the topics, thus, allowing for, both, the testing of the weakly or semi-supervised techniques presented in the SOTA, as well as the use of a more common evaluation metric, such as accuracy or F1 score.

As a result, 82 tweets were annotated according to a set of 13 predefined topics provided by MOG Technologies. Figure 43 presents the distribution of all the topics, which are the back pass, corner, foul, free kick, goal, kick off, offside, penalty, yellow card, red card, save, shot, and substitution.

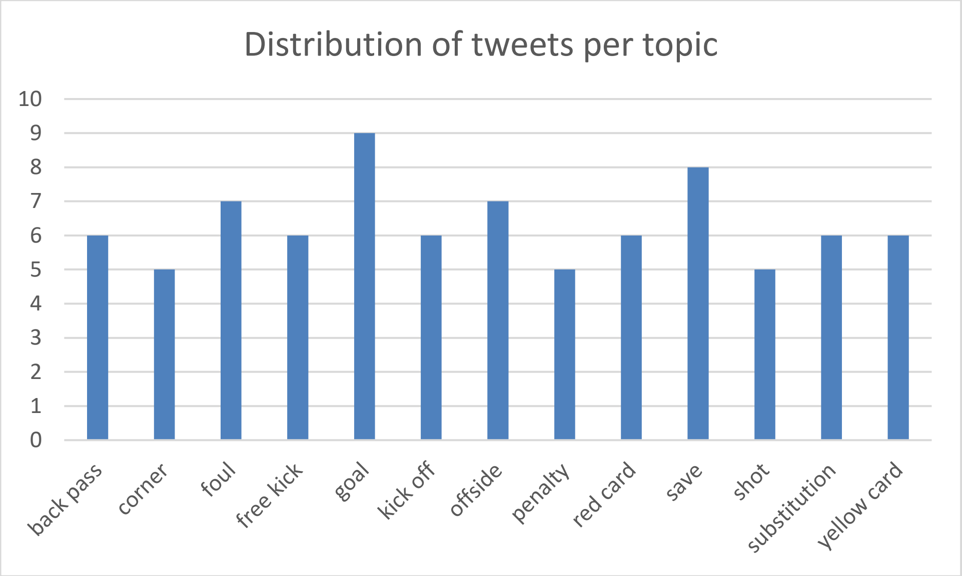


Figure 43 – Distribution of tweets per topic

4.3.2 Pre-processing experiments

With the set of topics predefined, the model to be used for carrying out the pre-processing experiments was chosen. Following a review of both weakly-supervised models presented in the preceding chapter’s Topic Analysis section, it was determined that the WeSHClass model was the best fit for the experiments, as an examination of the models' open-source code revealed that the JoSH model, due to its focus on topic discovery, lacked an implementation of its inference function.

As for the experiments’ evaluation, it was decided for the average pre-processing time per 100 instances, along with the weighted F1 score, as the distribution of the topic labels was unbalanced. In total, 8 experiments were carried out, using a sample of the dataset of 100.000 tweets for the training, with each either adding a new pre-processing simplification step to a previous experiment or replacing an already tested step.

As a result, the first experiment, in which all other experiments were based on, consisting in 11 pre-processing steps, was devised, taking into account the EDA of the dataset, the results of the

Processing experiments of the sentiment analysis task, and the pre-processing used in some of the papers already mentioned in the Text Mining chapter. This base experiment steps, ordered by application, consisted in the conversion of HTML characters into their respective ASCII form, the removal of handles, and hashtags, the replacement of URLs, resembles of laughs, emojis, and emoticons for their respective token, the removal of repeating characters, and accents, and the lowercasing of all text, as the WeSHClass model benefits of less complex text data for the generation of pseudo documents.

For the evaluation of the experiments, however, due to their randomness, each experiment was executed three times, so that the results' average, which are displayed in Figure 44, could be used as the comparison metric, as opposed to the results of single tests, which have a higher chance of being skewed.

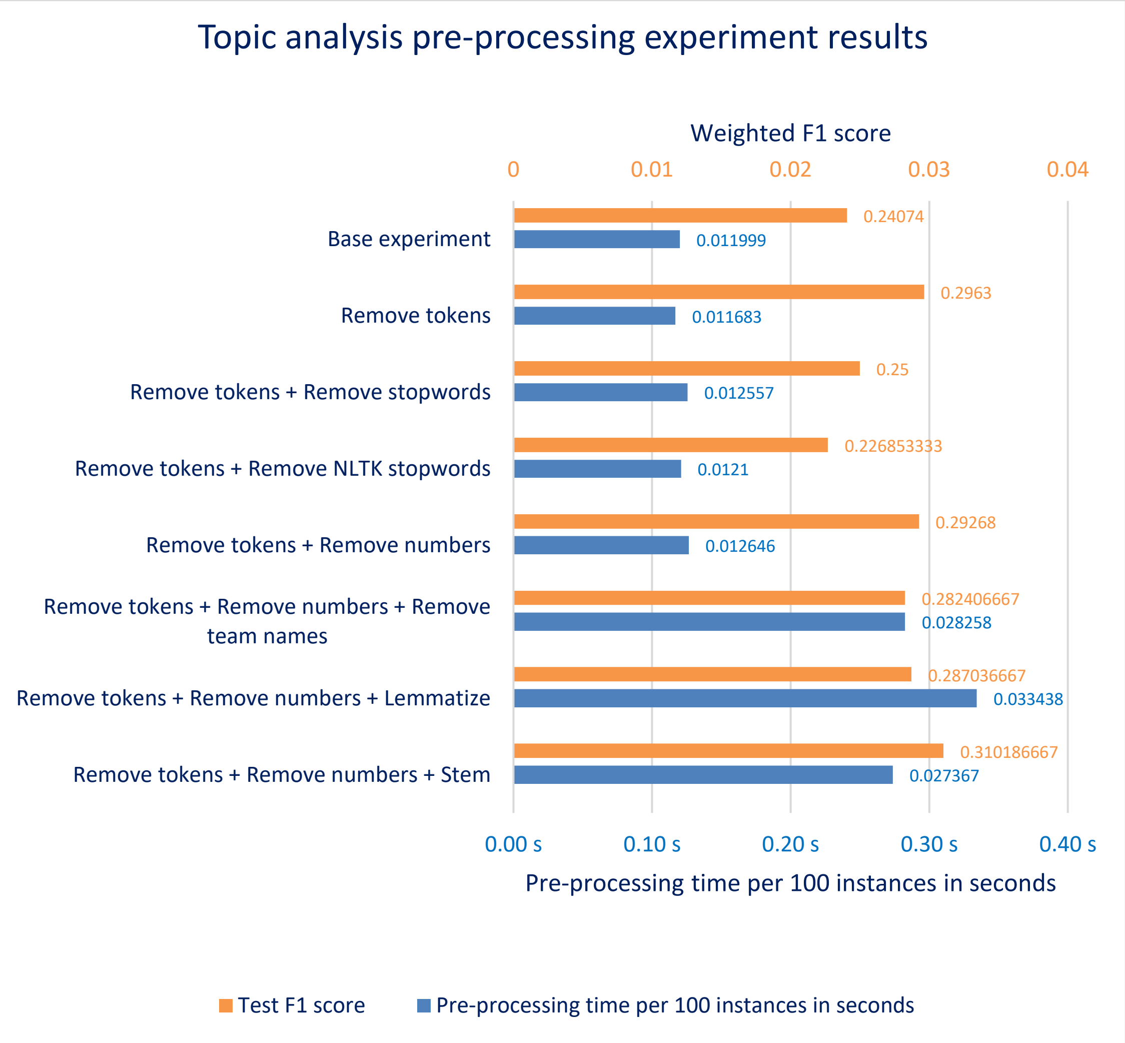


Figure 44 – Topic analysis pre-processing experiment results

Analysing the experiments' results, beforehand the selection of the best pre-processing approach, the inference time per 100 instances of the WeSHClass model was acquired, averaging at 0.7428 seconds across all the experiments. As such, accounting both metrics, the inference time, and the benefits of less complex text data for the generation of pseudo documents, the best pre-processing combination of steps was determined to be the one that build on the base experiment by removing all tokens and numbers, and stemming all words.

As a last note, it is important to mention that, following the same logic implemented in the pre-processing experiments of the sentiment analysis, the data formatting pre-processing time was included in the model inference time, rather than in the pre-processing time.

4.3.3 Processing experiments

With the pre-processing chosen, the WeSHClass model was then hyperparameter tuned using a set of experiments targeted at finding the optimal, or close to optimal, values for the maximization of its classification performance. As a result, 15 experiments were carried out, using a sample of 100.000 tweets, with the aim of helping with the selection of the best embeddings and pseudo document generation techniques, as well as the best, number of pre-training epochs and number of pseudo documents generated per topic.

All experiments were based on the default hyperparameters and techniques of the model, with each experiment iteratively changing only one of its values or techniques. As such, as presented in Figure 45, from the step-by-step changing of the first experiment, it was concluded that the pre-training of the model for five epochs on 500 pseudo documents generated using Skip-gram and an LSTM model achieved the best results.

As for the experiments' evaluation, following the same logic of the pre-processing experiments, each experiment was executed three times, thus allowing for the use of the average weighted F1 score as the comparison metric.

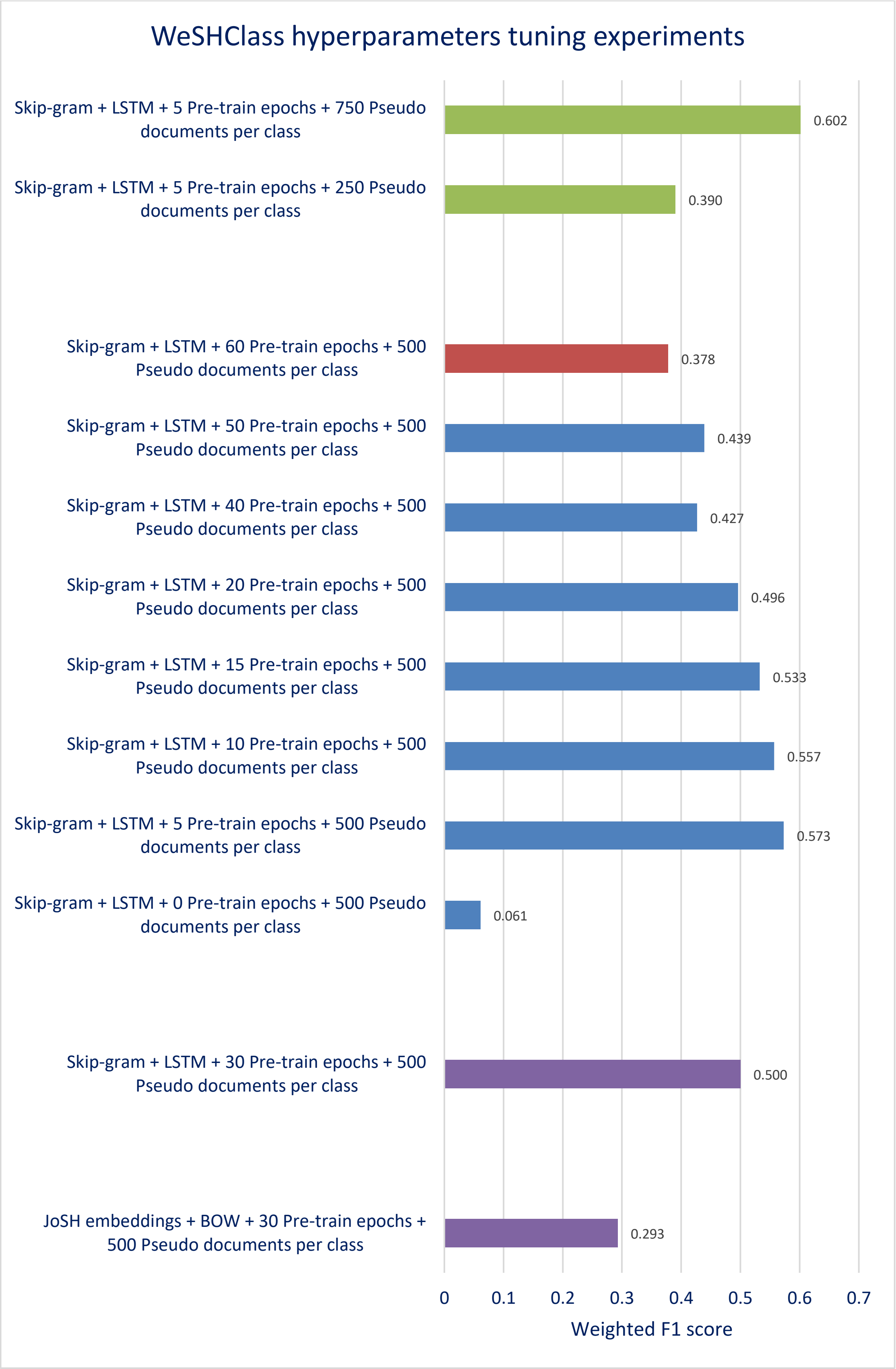


Figure 45 – WeSHClass hyperparameters tuning experiments

WeSHClass classification quality, however, does not only depend on the data and the hyperparameters, but also on the keywords provided for each topic. As such, in addition to the hyperparameter tuning experiments, it was tried several different combinations of keywords.

Table 7 presents three of the topics' keywords tested. With the own keywords, which were used for the hyperparameters tuning experiments, having been chosen through a combination of both, an analysis done on the tweets containing each topic word and the author knowledge of football. The JoSH keywords being the keywords, after removal of player and team's names, obtained from the JoSH model. And the improved keywords having been obtained from the continuous improvement of the own keywords, through the analysis of the confusion matrix, as well as the addition of keywords obtained from the JoSH topic mining results. As such, the improved keywords was able to achieve the average weighted F1 score of 0.492, thus proving to be better than, both, the own and JoSH keywords, which, respectively, only achieved F1 scores of 0.437 and 0.171.

Table 7 – WeSHClass topic keywords

Topics	Own keywords	JoSH keywords	Improved keywords
Back pass	back; pass	aerial; terrif; accuraci; motm; cap	back; pass
Corner	cross; pass	header; trick; cut; shoot; tackl	cross; pass; header
Foul	card; tackl	unnecessari; wide; behaviour; trick; oppon	tackl; dive
Free kick	foul; strike; cross	took; dream; save; gave; header	foul; strike; cross; header
Goal	score; assist	game; second; score; minut; assist	score; assist
Kick off	pm; matchday	ran; edg; threw; punch; blown	pm; matchday; blown
Offside	offsid; flag; var	rebound; deflect; elbow; shoot; counter	offsid; flag; var
Penalty	pen; penalti	ball; pen; shootout; bar; refere	pen; penalty; shootout
Yellow card	yellow; card	box; spot; yellow; cut; wall	yellow; card
Red card	red; card; var	wing; wank; across; allow; open	red; card; var; second
Save	goalkeep; keeper	lead; victori; dream; header; gave	goalkeep; keeper
Shot	strike; kick	space; feet; sane; situat; screw	strike; kick; space
Substitution	sub; substitut	queue; leg; broke; second; pull	sub; substitut

4.3.4 Final Model

Having the pre-processing techniques, the hyperparameter values, and the topics' keywords chosen, the model was, consequently, trained on the entire dataset, achieving a weighted F1 score of only 0.57, primarily due to its weak ability to correctly classify corner, save and shot related tweets, as seen in both the classification report and the confusion matrix presented in Figure 46.

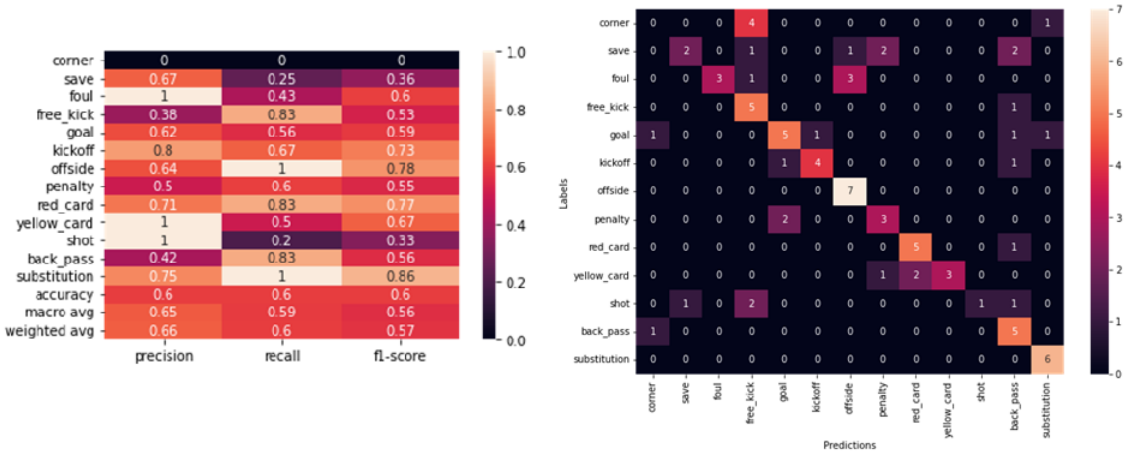


Figure 46 – WeSHClass 13 topics classification report and confusion matrix

As a result, for the improvement of the model performance, considering that shots are not a game defining moment, especially when the model already considers goals and saves, it was decided for its removal of the topics list. Additionally, to further help with the model performance, given the struggle of the model in differentiating corners from free kicks, and accounting for the fact that both can be considered standing still shots, it was decided for the junction of both topics into just the free kick topic. This, however, still allows for the distinction of both topics, as if over a period of time the tweets are relative to a free kick, it is more likely for their classification to be divided between free kick and foul, whereas if the tweets are relative to a corner, there is a lower chance of them being classified as foul.

With the topics' list now consisting in 11 topics, the training of the model on the entire dataset was redone, achieving a weighted F1 score of 0.61. Although, the increase in score was only of 0.04, upon a closer examination of the model results through its confusion matrix, depicted in Figure 47, it is possible to conclude that the model was able to improve in the classification of most topics, primarily having a low F1 score due to its difficulty in distinguishing between fouls, yellow and red cards related tweets, and save and penalty related tweets, which is understandable as these topics can overlap.

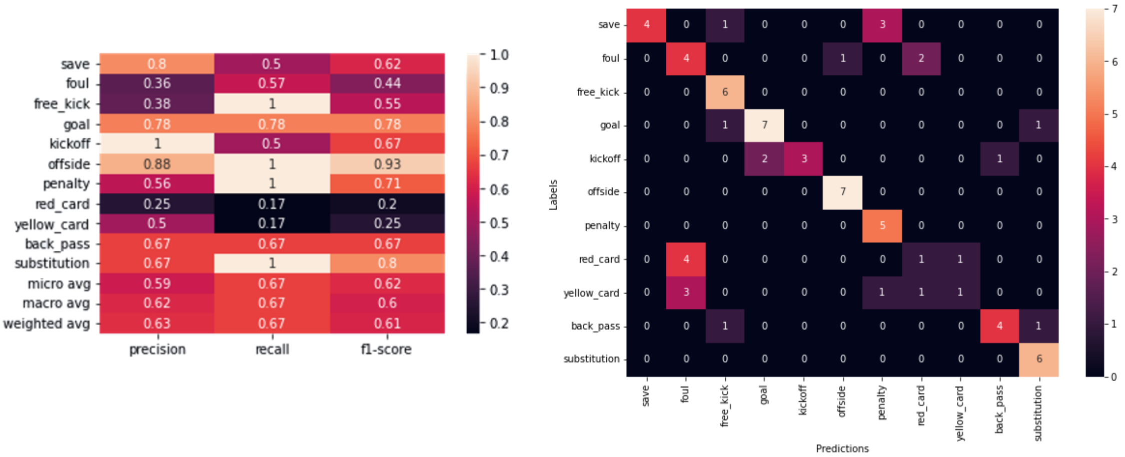


Figure 47 – WeSHClass 11 topics classification report and confusion matrix

4.4 Discussion of results

ML models and techniques, despite being increasingly more used in most areas, still suffer from the inability to fully explain their actions in a human understandable way. As such, when developing one of these tools, the process for improving its performance, still depends a lot on the execution of experiments, and consequently on the comparison of its results.

As such, for the development of the sentiment analysis tool, it was necessary for the execution of experiments for the selection of, both, the pre-processing techniques, and models. Regarding the pre-processing, it was concluded that oversimplifying the data harms the performance, thus leading to the pre-processing chosen consisting in the conversion of HTML characters into their respective ASCII form, the removal of handles and hashtags, the replacement of URLs, resembles of laughs, emojis and emoticons with tokens, and the removal of accents. Additionally, through these experiments it was also possible to conclude that emojis and emoticons, possible due to each having a different token, had a very minimal impact on the performance. As for the selection of the models, from the testing of the DistilBERT, DistilRoBERTa, and Knowledge embedding of both these models with teams' Wikipedia pages, it was concluded that the DistilRoBERTa with Knowledge embedding has the best performance. As a result, through the training of the knowledge embedded DistilRoBERTa model on the upsampled dataset pre-processed using the above-mentioned steps, it was achieved an accuracy of 94.7% with an average combined pre-processing and processing time of 0.1769 seconds per 100 instances.

As for the topic analysis, through an initial unsuccessful approach it was concluded that, both, the more traditional unsupervised statistical techniques, such as LDA, NMF, and LSA, and the coherence score evaluation metric are unsuitable, hence the decision for the partial annotation of the dataset regarding the topics, in order for the enabling of the use of the weakly-supervised hierarchical topic models, coupled with the F1 score. As such, using this metric for the pre-processing experiments, it was concluded that this task benefits more from less complex data than the sentiment analysis task, as the pre-processing that obtained the best result consists in the conversion of HTML characters into their respective ASCII form, the removal of handles, hashtags, URLs, resembles of laughs, emojis, emoticons, repeating characters, accents, and numbers, and the stemming and lowercasing of all words. Regarding the model selection, considering the lack of an inference function on the JoSH model, it was decided for the use of the WeSHClass hierarchical model, pre-trained for five epochs on 500 pseudo documents generated using Skip-gram and an LSTM model. As a result, through the training of the WeSHClass model on the dataset pre-processed using the above-mentioned steps, it was achieved a weighted F1 score of 0.61 on the classification of 11 topics, with an average combined pre-processing and processing time of 0.770167 seconds per 100 instances.

5 Conclusion

On this chapter, a summary of the developed work is presented, along with the conclusions reached from it. Additionally, the limitations of the sentiment and topic analysis tool are presented along with possible future development paths.

5.1 Summary and conclusions

Sport events' media consumption patterns have started transitioning to a multi-screen paradigm, where, through multitasking, viewers are able to search for additional information about the event they are watching live, as well as contribute with their perspective of the event to other viewers. The audiovisual and multimedia industries, however, are failing to capitalize on this by not providing the sports' teams and those in charge of the audiovisual production with insights on the final consumers perspective of sport events.

As a result of this opportunity, this document focuses on presenting the development of a near real-time sentiment analysis tool and a near real-time topic analysis tool for the analysis of sports events' related social media content that was published during the transmission of the respective events, thus enabling, in near real-time, the understanding of the sentiment of the viewers and the topics being discussed through each event.

For the development of both tools, it was first done a study on the social medias with the biggest potential to be used as the main data source, concluding on Twitter due to its availability of high-quality text content about most sport events, and its ability to stream the data rather than just provide it via GET requests. With the tools objectives defined and the data source chosen, an analysis on existing similar tools was done, concluding that none was capable of fully satisfying the intended objectives.

With the need for the development of a new tool, a study on the literature of, both, the fields of sentiment and topic analysis was done, concluding that, due to the best dataset found having sentiment labels for all its 6.3 million data points, but no topic labels, the DistilBERT, DistilRoBERTa, and Knowledge embedding of these pre-trained supervised models were the most promising methods for the sentiment analysis, and the JoSH and WeSHClass weakly-supervised models were the most promising methods for the topic analysis.

Prior to the testing and subsequently selection of the best processing technique for each task, due to the models' outputs being highly dependent on the quality of the data provided to them, it was first tested the best set of pre-processing steps. This not only allowed for the choosing of the most appropriate pre-processing for each task, but also allowed for the conclusion that the topic analysis task benefits more from less complex data than the sentiment analysis task. Additionally, through the experimentation of the several combinations of pre-processing steps, it was also concluded that the simple replacement of emojis and emoticons for their respective tokens has very minimal benefits for the sentiment analysis.

Having the pre-processing chosen, the sentiment analysis models were then tested, with the knowledge embedded DistilRoBERTa model achieving the best results. As such, through the training of this model on the upsampled dataset, simplified with the best combination of sentiment analysis pre-processing steps, it was achieved an accuracy of 94.7% with an average combined pre-processing and processing time of 0.1769 seconds per 100 instances.

Regarding the topic analysis, considering the lack of an inference function on the JoSH model, it was decided for the use of the WeSHClass model, pre-trained for five epochs on 500 pseudo documents generated using Skip-gram and an LSTM model. This model, trained on the dataset simplified with the best combination of topic analysis pre-processing steps, achieved a weighted F1 score of 0.61 on the classification of 11 topics, with an average combined pre-processing and processing time of 0. 770167 seconds per 100 instances. This F1 score although not very high, when examined through its confusion matrix, allowed for the conclusion that it had to do with similar topics overlapping, such as save and penalty or foul, red card, and yellow card, thus being misleading and making it seem that the topic analysis performance is worse than it really is.

Still regarding the topic analysis, through an initial unsuccessful approach it was also possible to conclude that, both, the more traditional unsupervised statistical techniques, such as LDA, NMF, and LSA, and the coherence score evaluation metric are unsuitable for the intended topic analysis tool, hence the decision for the partial annotation of the dataset regarding the topics, in order for the enabling of the use of the weakly-supervised hierarchical topic models, coupled with the F1 score.

In conclusion, in spite of both tools still having room for improvement, it is possible to affirm that the objectives of this dissertation were fulfilled and that this work is able to contribute to the scientific environment.

5.2 Future work

Regarding future work, in addition to the, already mentioned, support for more data sources, sports and languages, it was also identified several more possible future developments paths for, both, the sentiment and topic analysis tools.

As such, for the sentiment analysis tool, these paths consist in the:

- Grouping of emojis and emoticons, so that similar emojis and emoticons have the same token;
- Experimentation of knowledge embedding with different knowledge, such as, for example, previous games' scores;
- Testing of the final model on a manually annotated dataset;

For the topic analysis tool, these paths consist in the:

- Experimentation of different embeddings and pseudo-document generation techniques for the pre-training of the WeSHClass model;
- Improvement of hyperparameter fine-tuning for the WeSHClass model using smaller value intervals between experiments;
- Improvement of the topics and keywords provided to the WeSHClass, with the help of more topic mining models, as well as by analysing if the WeSHClass model is blocking classifications from reaching the final model, which means that no topic is representative of the text classified;
- Annotation of more data in regard to the topics, in order to improve the test set size;

Additionally, as future work it is also planned the writing of a scientific paper, to be submitted to, either a journal or conference, regarding the work done for both the sentiment and topic analysis tools.

References

Abd El-Jawad, M. H., Hodhod, R., & Omar, Y. M. K. (2019). Sentiment analysis of social media networks using machine learning. *ICENCO 2018 - 14th International Computer Engineering Conference: Secure Smart Societies*, 174–176. https://doi.org/10.1109/ICENCO.2018.8636124

Aggarwal, C. C., & Zhai, C. X. (2013). Mining text data. In *Mining Text Data* (Vol. 9781461432). https://doi.org/10.1007/978-1-4614-3223-4

Ahmad, M., Aftab, S., & Ali, I. (2017). Sentiment Analysis of Tweets using SVM. *International Journal of Computer Applications*, *177*(5), 25–29. https://doi.org/10.5120/ijca2017915758

Ahmad, M., Aftab, S., Muhammad, S., & Ahmad, S. (2017). Machine Learning Techniques for Sentiment Analysis: A Review. *Int. J. Multidiscip. Sci. Eng*, *8*(3), 27–32. www.ijmse.org

Akkaya, C., Wiebe, J., & Mihalcea, R. (2009). Subjectivity word sense disambiguation. *EMNLP 2009 - Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: A Meeting of SIGDAT, a Special Interest Group of ACL, Held in Conjunction with ACL-IJCNLP 2009*, 190–199. https://doi.org/10.3115/1699510.1699535

Alake, R. (2020). *Understanding Cosine Similarity And Its Application*. Towards Data Science. https://towardsdatascience.com/understanding-cosine-similarity-and-its-application-fd42f585296a

Alammar, J. (2021). *The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning) – Jay Alammar – Visualizing machine learning one concept at a time.* 1–19. http://jalammar.github.io/illustrated-bert/

An, S., Ji, L. J., Marks, M., & Zhang, Z. (2017). Two sides of emotion: Exploring positivity and negativity in six basic emotions across cultures. In *Frontiers in Psychology* (Vol. 8, Issue SEP, p. 610). Frontiers Media S.A. https://doi.org/10.3389/fpsyg.2017.01467

Araque, O., Zhu, G., & Iglesias, C. A. (2019). A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowledge-Based Systems*, *165*, 346–359. https://doi.org/10.1016/j.knosys.2018.12.005

Asadollahfardi, G. (2015). Artificial Neural Network. In *Interdisciplinary Computing in Java Programming* (pp. 77–91). Springer, Boston, MA. https://doi.org/10.1007/978-3-662-44725-3_5

Ayvaz, S., & Shiha, M. O. (2017). The Effects of Emoji in Sentiment Analysis. *International Journal of Computer and Electrical Engineering*, *9*(1), 360–369. https://doi.org/10.17706/ijcee.2017.9.1.360-369

Barnes, J. A. (1954). Class and Committees in a Norwegian Island Parish. *Human Relations*, *7*(1), 39–58. https://doi.org/10.1177/001872675400700102

BetSentiment. (2019). *Sentiment Analysis | Premier League Players, Teams & Predictions | Betsentiment.com*. https://betsentiment.com/

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(4–5), 993–1022. https://doi.org/10.1016/b978-0-12-411519-4.00006-9

Bonta, V., Kumaresh, N., & Janardhan, N. (2019). A Comprehensive Study on Lexicon Based Approaches for Sentiment Analysis. *Asian Journal of Computer Science and Technology*, *8*(S2), 1–6. https://doi.org/10.51983/ajcst-2019.8.s2.2037

Bouma, G. (2009). Normalized (Pointwise) Mutual Information in Collocation Extraction. *Proceedings of German Society for Computational Linguistics (GSCL 2009)*, 31–40.

Bradley, M. M., & Lang, P. J. (2008). Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings. *IEEE Internet Computing*, *12*(5), 44–52. http://dionysus.psych.wisc.edu/methods/Stim/ANEW/ANEW.pdf%5Cnhttp://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Affective+Norms+for+English+Words+(+ANEW+):+Instruction+Manual+and+Affective+Ratings#0%5Cnhttp://scholar.google.com/scholar?hl=en&bt

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *2020-Decem*.

https://arxiv.org/abs/2005.14165v4

Campos, J. R. P., Gamallo, P., & Alegria, I. (2018). Measuring language distance among historical varieties using perplexity. Application to European Portuguese. *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects*, 145–155.

Carr, C. T., & Hayes, R. A. (2015). Social Media: Defining, Developing, and Divining. *Atlantic Journal of Communication*, *23*(1), 46–65. https://doi.org/10.1080/15456870.2015.972282

ChannelMeter. (2019). *YouTube's Top Countries*. Medium. https://medium.com/@ChannelMeter/youtubes-top-countries-47b0d26dded#

Codd, E. F. (1970). A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM*, *13*(6), 377–387. https://doi.org/https://doi.org/10.1145/362384.362685

Coelho, T. da S. (2021). *Análise de publicações no Twitter nas Instituições do Ensino Superior do topo de Ranking Mundial*. https://hdl.handle.net/10216/138531

Crisci, A., Grasso, V., Nesi, P., Pantaleo, G., Paoli, I., & Zaza, I. (2018). Predicting TV programme audience by using twitter based metrics. *Multimedia Tools and Applications*, *77*(10), 12203–12232. https://doi.org/10.1007/s11042-017-4880-x

Crocetti, P., Peterson, S., & Hefner, K. (2021, February). *What is Data Protection and Why is it Important? Definition from WhatIs.com*. TechTarget. https://www.techtarget.com/searchdatabackup/definition/data-protection

Curry, D. (2021). *Discord Revenue and Usage Statistics (2021)*. Business of Apps. https://www.businessofapps.com/data/discord-statistics/

Daily, S. B., James, M. T., Cherry, D., Porter, J. J., Darnell, S. S., Isaac, J., & Roy, T. (2017). Affective Computing: Historical Foundations, Current Applications, and Future Trends. In *Emotions and Affect in Human Factors and Human-Computer Interaction* (pp. 213–231). Academic Press. https://doi.org/10.1016/B978-0-12-801851-4.00009-4

Dalianis, H. (2018). Clinical text mining: Secondary use of electronic patient records. In *Clinical Text Mining: Secondary Use of Electronic Patient Records*. https://doi.org/10.1007/978-3-319-78503-5

Darwich, M., Mohd Noah, S. A., Omar, N., & Osman, N. A. (2019). Corpus-Based Techniques for Sentiment Lexicon Generation: A Review. *Journal of Digital Information Management*, *17*(5), 296. https://doi.org/10.6025/jdim/2019/17/5/296-305

Dean, B. (2021). *Twitch Usage and Growth Statistics: How Many People Use Twitch in 2021?* Backlinko. https://backlinko.com/twitch-users

DeepAI. (2020). *Evaluation Metrics Definition | DeepAI*. DeepAI. https://deepai.org/machine-learning-glossary-and-terms/evaluation-metrics

Degenhard, J. (2021). *Reddit user worldwide 2020, by country*. Statista. https://www.statista.com/forecasts/1174696/reddit-user-by-country

Delaporte, A., & Bahia, K. (2021). The State of Mobile Internet Connectivity 2021. In *GSMA Reports*. www.gsmaintelligence.com

Deliciousavocado. (2021). *youtube - How much quota cost does the LiveChatMessages.list method incur? - Stack Overflow*. Stack Overflow. https://stackoverflow.com/questions/67232262/how-much-quota-cost-does-the-livechatmessages-list-method-incur.

DeVito, J. A., O'Rourke, S., & O'Neill, L. (2000). *Human communication*. Longman New York.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, *1*, 4171–4186. https://arxiv.org/abs/1810.04805v2

Dhruvil Karani. (2018). *Introduction to Word Embedding and Word2Vec*. Towars Data Science. https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE*, *6*(12), e26752. https://doi.org/10.1371/journal.pone.0026752

Dumais, S. T. (2004). Latent Semantic Analysis. In *Annual Review of Information Science and Technology* (Vol. 38, Issue 1, pp. 188–230). John Wiley & Sons, Ltd. https://doi.org/10.1002/aris.1440380105

Esuli, A., & Sebastiani, F. (2006). SENTIWORDNET: A publicly available lexical resource for opinion mining. *Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006*, 417–422. http://www-2.cs.cmu.edu/

EzBUCKETz. (2021). *Los Angeles Lakers vs Detroit Pistons NBA LIVE Play-By-Play & Reaction*. YouTube. https://www.youtube.com/watch?v=C4UUPb-y1u8

Facebook. (2021a). *Facebook - Facebook Reports Third Quarter 2021 Results*. Facebook. https://investor.fb.com/investor-news/press-release-details/2021/Facebook-Reports-Third-Quarter-2021-Results/default.aspx

Facebook. (2021b). *What's the difference between a profile, Page and group on Facebook?* Facebook. https://www.facebook.com/help/337881706729661

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge Discovery and Data Mining: Towards a Unifying Framework. *Int Conf on Knowledge Discovery and Data Mining*, 82–88. https://www.aaai.org/Papers/KDD/1996/KDD96-014.pdf

Feldman, R., & Dagan, I. (1995). Knowledge Discovery in Textual Databases (KDT). *International Conference on Knowledge Discovery and Data Mining (KDD)*, 112–117. www.aaai.org

Fitelson, B. (2003). A probabilistic theory of coherence. *Analysis*, *63*(3), 194–199. https://doi.org/10.1093/analys/63.3.194

Garbe, W. (2015). *Fast approximate string matching with large edit distances in Big Data*. Medium. https://wolfgarbe.medium.com/fast-approximate-string-matching-with-large-edit-distances-in-big-data-2015-9174a0968c0b

Gligorić, K., Anderson, A., & West, R. (2020). *Adoption of Twitter's New Length Limit: Is 280 the New 140?* https://arxiv.org/abs/2009.07661v1

GMI Blogger. (2021). *YOUTUBE USER STATISTICS 2021*. Global Media Insight. https://www.globalmediainsight.com/blog/youtube-users-statistics/%0A

Google Developers. (2020a). *Classification: Accuracy | Machine Learning Crash Course | Google Developers*. Machine Learning Crash Course. https://developers.google.com/machine-learning/crash-course/classification/accuracy

Google Developers. (2020b). *Classification: Precision and Recall | Machine Learning Crash Course | Google Developers*. Machine Learning Crash Course. https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall

Google Developers. (2020c). *Classification: True vs. False and Positive vs. Negative | Machine Learning Crash Course | Google Developers*. Machine Learning Crash Course. https://developers.google.com/machine-learning/crash-course/classification/true-false-positive-negative

Google, & Ipsos Connect. (2017). *TV viewership behavior data - Think with Google*. Think with Google. https://www.thinkwithgoogle.com/consumer-insights/consumer-trends/sports-fans-video-insights-6/

Gupta, A. (2021). *A Comprehensive Guide on Deep Learning Optimizers*. Data Science Blogathon. https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-deep-learning-optimizers/

Gupta, N., & Agrawal, R. (2020). Application and techniques of opinion mining. In *Hybrid Computational Intelligence* (pp. 1–23). Academic Press. https://doi.org/10.1016/b978-0-12-818699-2.00001-9

Gutiérrez, L., Bekios-Calfa, J., & Keith, B. (2019). A review on bayesian networks for sentiment analysis. *Advances in Intelligent Systems and Computing*, *865*, 111–120. https://doi.org/10.1007/978-3-030-01171-0_10

Guynn, J. (2018). *Facebook news feed: The giant social network is making a big change*. USA TODAY. https://eu.usatoday.com/story/tech/2018/01/11/facebook-newsfeed-big-change/1023331001/

Han, J., Kamber, M., & Pei, J. (2012). Data Mining Trends and Research Frontiers. In *Data Mining* (pp. 585–631). Morgan Kaufmann. https://doi.org/10.1016/b978-0-12-381479-1.00013-7

Hartman, J. J., Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvia, D. M. (1967). The General Inquirer: A Computer Approach to Content Analysis. *American Sociological Review*, *32*(5), 859. https://doi.org/10.2307/2092070

Healey, & Ramaswamy. (2013). *Twitter Sentiment Visualization*. https://www.csc2.ncsu.edu/faculty/healey/tweet_viz/

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. *LDV Forum - GLDV Journal* *for Computational Linguistics and Language Technology*, *20*, 19–62. https://doi.org/10.1111/j.1365-2621.1978.tb09773.x

Huang, S., Niu, Z., & Shi, C. (2014). Automatic construction of domain-specific sentiment lexicon based on constrained label propagation. *Knowledge-Based Systems*, *56*, 191–200. https://doi.org/10.1016/j.knosys.2013.11.009

Hugging Face. (2019). *distilroberta-base · Hugging Face*. https://huggingface.co/distilroberta-base

Huilgol, P. (2019). *Accuracy vs. F1-Score*. Analytics Vidhya. https://medium.com/analytics-vidhya/accuracy-vs-f1-score-6258237beca2

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, 216–225.

IBM. (2021). *Structured vs. Unstructured Data: What's the Difference? | IBM*. IBM Information. https://www.ibm.com/cloud/blog/structured-vs-unstructured-data

Iyer, G., Soberman, D., & Villas-Boas, J. M. (2005). The targeting of advertising. In *Marketing Science* (Vol. 24, Issue 3, pp. 461–476). INFORMS. https://doi.org/10.1287/mksc.1050.0117

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: A review. *ACM Computing Surveys*, *31*(3), 264–323. https://doi.org/10.1145/331499.331504

Jiang, H., He, P., Chen, W., Liu, X., Gao, J., & Zhao, T. (2020). SMART: Robust and Efficient Fine-Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2177–2190. https://doi.org/10.18653/v1/2020.acl-main.197

Jovanoski, D., Pachovski, V., & Nakov, P. (2016). On the impact of seed words on sentiment polarity lexicon induction. *COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers*, 1557–1567.

K., J., & R., J. (2016). Stop-Word Removal Algorithm and its Implementation for Sanskrit Language. *International Journal of Computer Applications*, *150*(2), 15–17. https://doi.org/10.5120/ijca2016911462

Kapadia, S. (2019). *Evaluate Topic Models: Latent Dirichlet Allocation (LDA)*. Towards Data Science. https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis* (L. Kaufman & P. J. Rousseeuw (eds.)). John Wiley & Sons, Inc. https://doi.org/10.1201/9780429470615-4

Kemp, S. (2021a). *Digital 2021: Global Overview Report — DataReportal – Global Digital Insights*. Kepios Pte. Ltd., We Are Social Ltd. and Hootsuite Inc. https://datareportal.com/reports/digital-2021-global-overview-report

Kemp, S. (2021b). *Digital 2021 October Global Statshot Report — DataReportal – Global Digital Insights*. Kepios. https://datareportal.com/reports/digital-2021-october-global-statshot

Kemp, S. (2021c). *Essential Twitter stats for 2021*. Kepios. https://datareportal.com/essential-twitter-stats

Kesiraju, S., Plchot, O., Burget, L., & Gangashetty, S. V. (2020). Learning Document Embeddings along with Their Uncertainties. *IEEE/ACM Transactions on Audio Speech and Language Processing*, *28*, 2319–2332. https://doi.org/10.1109/TASLP.2020.3012062

Kherwa, P., & Bansal, P. (2018). Topic Modeling: A Comprehensive Review. *ICST Transactions on Scalable Information Systems*, *0*(0), 16. https://doi.org/10.4108/eai.13-7-2018.159623

Kim, S. M., & Hovy, E. (2004). Determining the sentiment of opinions. *COLING 2004 - Proceedings of the 20th International Conference on Computational Linguistics*. https://doi.org/10.3115/1220355.1220555

Kingma, D. P., & Ba, J. L. (2015, December 22). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. https://doi.org/10.48550/arxiv.1412.6980

Krause, B., Murray, I., Renals, S., & Lu, L. (2019, September 26). Multiplicative LSTM for sequence modelling. *5th International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings*. https://arxiv.org/abs/1609.07959v3

Krüger, F. (2018). Activity, Context, and Plan Recognition with Computational Causal Behaviour Models.

ResearchGate, *August*. https://www.researchgate.net/publication/314116591_Activity_Context_and_Plan_Recognition_with_Computational_Causal_Behaviour_Models

Kumar, K. (2018). *Evaluation of Topic Modeling: Topic Coherence*. Data Science +. https://datascienceplus.com/evaluation-of-topic-modeling-topic-coherence/

Lai, G., Xie, Q., Liu, H., Yang, Y., & Hovy, E. (2017). RACE: Large-scale ReAding comprehension dataset from examinations. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 785–794. https://doi.org/10.18653/v1/d17-1082

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791. https://doi.org/10.1038/44565

Lee, H., & Bhd, C. ePulze S. (2011). Chinese Sentiment Analysis Using Maximum Entropy. *Sentiment Analysis …*, *72*, 89–93. http://acl.eldoc.ub.rug.nl/mirror/W/W11/W11-37.pdf#page=105

Lerer, A., Wu, L., Shen, J., Lacroix, T., Wehrstedt, L., Bose, A., & Peysakhovich, A. (2019). *PyTorch-BigGraph: A Large-scale Graph Embedding System*. https://github.com/facebookresearch/

Li, T. (2019). *Argmax vs Softmax vs Sparsemax - Tao Li*. https://www.cs.utah.edu/~tli/posts/2019/01/blog-post-1/

Liu, W., Zhou, P., Zhao, Z., Wang, Z., Ju, Q., Deng, H., & Wang, P. (2020). K-BERT: Enabling language representation with knowledge graph. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, *34*(03), 2901–2908. https://doi.org/10.1609/aaai.v34i03.5681

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. https://arxiv.org/abs/1907.11692v1

Lutz, C., & White, G. M. (1986). The Anthropology of Emotions. *Annual Review of Anthropology*, *15*(1), 405–436. https://doi.org/10.1146/annurev.an.15.100186.002201

Ma, B., Yuan, H., & Wu, Y. (2017). Exploring performance of clustering methods on document sentiment analysis. *Journal of Information Science*, *43*(1), 54–74. https://doi.org/10.1177/0165551515617374

Manning, C. D., Raghavan, P., & Schütze, H. (2009). Introduction to information retrieval. *Choice Reviews Online*, *46*(05), 46-2715-46–2715. https://doi.org/10.5860/choice.46-2715

Mannor, S., Jin, X., Han, J., Jin, X., Han, J., Jin, X., Han, J., & Zhang, X. (2011). K-Means Clustering. In *Encyclopedia of Machine Learning* (pp. 563–564). Springer, Boston, MA. https://doi.org/10.1007/978-0-387-30164-8_425

Mardia, K. V. (1975). Distribution Theory for the Von Mises-Fisher Distribution and Its Application. In *A Modern Course on Statistical Distributions in Scientific Work* (pp. 113–130). Springer, Dordrecht. https://doi.org/10.1007/978-94-010-1842-5_10

Mark Goldbridge That's Football. (2021). *ENGLAND vs ITALY LIVE EURO 202O Final Watchalong Mark GOLDBRIDGE LIVE*. YouTube. https://www.youtube.com/watch?v=5ef6g7b2G9U

Mayfield, III, T. D. (2011). A Commander's Strategy for Social Media. *Joint Force Quarterly*, *1st quarte*(60), 79–83. http://www.au.af.mil/au/awc/awcgate/jfq/mayfield_strat_for_soc_media.pdf

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Meng, Y., Shen, J., Zhang, C., & Han, J. (2019). Weakly-supervised hierarchical text classification. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, 6826–6833. https://doi.org/10.1609/aaai.v33i01.33016826

Meng, Y., Zhang, Y., Huang, J., Zheng, Y., Zhang, C., & Han, J. (2020). Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1908–1917. https://doi.org/10.1145/3394486.3403242

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013, January 16). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*. https://doi.org/10.48550/arxiv.1301.3781

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748

Mitchell, J. C. (1974). Social Networks. *Annual Review of Anthropology*, *3*(1), 279–299.

https://doi.org/10.1146/annurev.an.03.100174.001431

Moody, C. E. (2016). *Mixing Dirichlet Topic Models and Word Embeddings to Make lda2vec*. https://arxiv.org/abs/1605.02019v1

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*.

Niwattanakul, S., Singthongchai, J., Naenudorn, E., & Wanapu, S. (2013). Using of Jaccard Coefficient for Keywords Similarity Effectiveness of Constrained Handling Techniques of Improved Constrained Differential Evolution Algorithm Applied to Constrained Optimization Problems in Mechanical Engineering View project Data mining. *Proceedings of the International MultiConference of Engineers and Computer Scientists*, *I*. https://www.researchgate.net/publication/317248581

Osgood, C. E., May, W. H., & Miron, M. S. (1975). *Cross-Cultural Universals of Affective Meaning* (Vol. 1). University of Illinois Press. https://books.google.pt/books/about/Cross_cultural_Universals_of_Affective_M.html?id=Ax9CFVxXYD4C&redir_esc=y

Ostendorff, M., Bourgonje, P., Berger, M., Moreno-Schneider, J., Rehm, G., & Gipp, B. (2019). Enriching BERT with knowledge graph embeddings for document classification. *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019*, 307–314. https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/germeval/Germeval_Task1_paper_3.pdf

Parveen, H., & Pandey, S. (2017). Sentiment analysis on Twitter Data-set using Naive Bayes algorithm. *Proceedings of the 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology, ICATccT 2016*, 416–419. https://doi.org/10.1109/ICATCCT.2016.7912034

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720

Picard, R. W., Papert, S., Bender, W., Blumberg, B., Breazeal, C., Cavallo, D., Machover, T., Resnick, M., Roy, D., & Strohecker, C. (2004). Affective learning - a manifesto. *BT Technology Journal*, *22*(4), 253–269. https://doi.org/10.1023/B:BTTJ.0000047603.37042.33

Poushter, J., Bishop, C., & Chwe, H. (2018). Social media use continues to rise in developing countries but plateaus across developed ones: smartphone ownership on the rise in emering economies. In *Pew Research Center* (Vol. 19). www.pewresearch.org.

Radford, A., Jozefowicz, R., & Sutskever, I. (2017). *Learning to Generate Reviews and Discovering Sentiment*. https://arxiv.org/abs/1704.01444v2

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuad: 100,000+ questions for machine comprehension of text. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2383–2392. https://doi.org/10.18653/v1/d16-1264

Rajput, R., & Solanki, A. K. (2016). Review of Sentimental Analysis Methods using Lexicon Based Approach. *International Journal of Computer Science and Mobile Computing*, *5*(2), 159–166. www.ijcsmc.com

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. https://arxiv.org/abs/1910.01108v4

Saravanakumar, M., & SuganthaLakshmi, T. (2012). Social media marketing. *Life Science Journal*, *9*(4), 4444–4451. https://doi.org/10.5937/markt1704254k

Sasaki, Y. (2007). *The truth of the F-measure* (Issue January 2007).

Schneider, J. (1997). *Cross Validation*. https://www.cs.cmu.edu/~schneide/tut5/node42.html

Scott, J. (2002). *Social Networks: Critical Concepts in Sociology* (Vol. 4). Taylor & Francis. https://books.google.pt/books?hl=pt-PT&lr=&id=u1le8gcwTcsC&oi=fnd&pg=PR4&dq=social+networks+a+sociology+review&ots=eR39psm9nI&sig=FSJKrTz4s1m1ZACE8KzIJkmvH-8&redir_esc=y#v=onepage&q=social networks a sociology review&f=false

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 888–905. https://doi.org/10.1109/34.868688

Sporting Clube de Portugal. (2021). *Equipa B | FC Ol. Hospital x Sporting CP (Jogo Completo)*. YouTube. https://www.youtube.com/watch?v=2naLMVr7kmw

Statista. (2021). *U.S. Reddit app users by age 2021*. Statista. https://www.statista.com/statistics/1125159/reddit-us-app-users-age/

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, *62*(1), 77–89. https://doi.org/10.1016/S0034-4257(97)00083-7

Sterley, T. L., & Bains, J. S. (2021). Social communication of affective states. In *Current Opinion in Neurobiology* (Vol. 68, pp. 44–51). Elsevier Current Trends. https://doi.org/10.1016/j.conb.2020.12.007

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*(2), 267–307. https://doi.org/10.1162/COLI_a_00049

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. In *Journal of Language and Social Psychology* (Vol. 29, Issue 1, pp. 24–54). SAGE PublicationsSage CA: Los Angeles, CA. https://doi.org/10.1177/0261927X09351676

Tul, Q., Ali, M., Riaz, A., Noureen, A., Kamranz, M., Hayat, B., & Rehman, A. (2017). Sentiment Analysis Using Deep Learning Techniques: A Review. *International Journal of Advanced Computer Science and Applications*, *8*(6). https://doi.org/10.14569/ijacsa.2017.080657

Turing, A. M. (1950). Computer Machinery and Intelligence. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/MIND/LIX.236.433

Turney, P. D. (2002). *Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews*. https://doi.org/10.3115/1073083.1073153

Twitter Data. (2014). *Twitter Data on Twitter: "With 35.6 million Tweets, #BRA v #GER is the most-discussed single sports game ever on Twitter. #WorldCup http://t.co/pRjssAZmhg" / Twitter*. Twitter. https://twitter.com/twitterdata/status/486708145775841281

Twitter Developer Platform. (2021). *Twitter API Documentation | Docs | Twitter Developer Platform*. Twitter Developer Platform. https://developer.twitter.com/en/docs/twitter-api

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *2017-Decem*, 5999–6009.

Velikovich, L., Blair-Goldensohn, S., Hannan, K., & McDonald, R. (2010). The viability of web-derived polarity lexicons. *NAACL HLT 2010 - Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings of the Main Conference*, 777–785.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *17*(4), 395–416. https://doi.org/10.1007/s11222-007-9033-z

Voorveld, H. A. M., van Noort, G., Muntinga, D. G., & Bronner, F. (2018). Engagement with Social Media and Social Media Advertising: The Differentiating Role of Platform Type. *Journal of Advertising*, *47*(1), 38–54. https://doi.org/10.1080/00913367.2017.1405754

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). Glue: A multi-task benchmark and analysis platform for natural language understanding. *7th International Conference on Learning Representations, ICLR 2019*. https://doi.org/10.18653/v1/w18-5446

Wang, S., Fang, H., Khabsa, M., Mao, H., & Ma, H. (2021). *Entailment as Few-Shot Learner*. http://arxiv.org/abs/2104.14690

Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., & Tang, J. (2021). Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, *9*, 176–194. https://doi.org/10.1162/TACL_A_00360/98089/KEPLER-A-UNIFIED-MODEL-FOR-KNOWLEDGE-EMBEDDING-AND

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*(4), 1191–1207. https://doi.org/10.3758/s13428-012-0314-x

We Are Social; DataReportal; Hootsuite. (2021, January). *Daily social media usage worldwide | Statista*. Statista. https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/

Xie, Q., Dai, Z., Hovy, E., Luong, M. T., & Le, Q. V. (2020). Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, *2020-Decem*.

https://github.com/google-research/uda.

Yan, X., Guo, J., Liu, S., Cheng, X., & Wang, Y. (2013). Learning topics in short texts by non-negative matrix factorization on term correlation matrix. *Proceedings of the 2013 SIAM International Conference on Data Mining, SDM 2013*, 749–757. https://doi.org/10.1137/1.9781611972832.83

Yao, L., Mimno, D., & McCallum, A. (2009). Efficient methods for topic model inference on streaming document collections. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 937–945. https://doi.org/10.1145/1557019.1557121

Yi Li, K. (2021). *How to Choose a Learning Rate Scheduler for Neural Networks*. Neptune AI. https://neptune.ai/blog/how-to-choose-a-learning-rate-scheduler

Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). *Comparative Study of CNN and RNN for Natural Language Processing*. https://arxiv.org/abs/1702.01923v1

Yuan, J., Gao, F., Ho, Q., Dai, W., Wei, J., Zheng, X., Xing, E. P., Liu, T. Y., & Ma, W. Y. (2015). LightLDA: Big topic models on modest computer clusters. *WWW 2015 - Proceedings of the 24th International Conference on World Wide Web*, 1351–1361. https://doi.org/10.1145/2736277.2741115

Zhao, Y., & Karypis, G. (2005). Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, *10*(2), 141–168. https://doi.org/10.1007/s10618-005-0361-3