COMP61332 Text Mining Coursework 2: Social Media Analytics

Groups 16 & 26

Darsh Jadhav ID: *10806664*

Huajie He ID: *10689523*

Jonathan Coutinho ID: *9931026*

Nuhu Ibrahim ID: *10723572*

Xinmiao Tang ID: *10618359*

Yuqi Zhang ID: *10613157*

***Abstract*—“Social Media Analytics” (*SMA*) is a flourishing field that makes it possible to reap the benefits of large amounts of data generated daily from billions of social media users worldwide. It is a way to understand the “word on the street” or “scoop” about popular subjects. This paper demonstrates our implementation of the *SMA* pipeline on data related to the 2020 Summer Olympic Games (Tokyo 2020 or Tokyo 202One). Postponed due to *COVID-19*, the games will take place in July 2021 and are set to make highlights as one of the year’s first worldwide in-person public events. Consequently, there has been lots of controversy floating around social media, thus we have decided to investigate specific event-related themes to evaluate the public’s opinions. Our findings indicate a significant increase in *negative* emotion over the course of *October 2019-March 2021*. Furthermore, we determined the “*hottest topics*” surrounding the event. The applications of our research are beneficial to any and all beings associated with the Olympic games, from people looking to learn more about the Olympics, to businesses attempting to increase sales by using the Olympics to their advantage.**

***Index Terms*—Social Media Analytics, Tokyo Olympics, Twitter, Topic Modelling, Sentiment Analysis, Named Entity Recognition, Named Entity Linking**

I. Introduction

In 2018, Japan won the rights to host the 2020 Olympic games, i.e. “Tokyo 2020”. This was a significant milestone for Japan, and was set to be one of the most anticipated events of the year [1]. However, the emergence of an unfounded, evolving virus known as “Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV2)” [2] played spoiler. There was a severe outbreak of the novel coronavirus disease 2019 (COVID-19), which resulted in the closure and cease of all public places and events across the world. In *March 2020*, former Japanese Prime Minister *Shinzo Abe*, and President of the International Olympic Committee (IOC), *Thomas Bach* announced the postponement of Tokyo 2020 to 2021 [3], [4].

Lauded as the world’s biggest international multi-sport event [5], there is now a lot of hope that the games will serve as an avenue to the celebration of more large-scale public events [3]. This paper aims to conduct an analysis around the upcoming event using social media, specifically *Twitter*. Using a modified version of developed frameworks [6]–[8], we propose mining for data around the following themes:

1) What topics are discussed by people regarding the Tokyo Olympics? [**Topic Modelling**]
2) What are the major sentiments about the Tokyo Olympics? [**Sentiment Analysis**]
3) Have peoples’ opinions on the Tokyo Olympics changed over the period of COVID-19? [**Sentiment Analysis**]
4) Are people scared that there will be a surge in COVID-19 cases in relation to the 2020 Summer Olympics? [**Sentiment Analysis**]
5) What specific organizations, people, events, or places are the most talked about in relation to the Tokyo Olympics? [**Named Entity Recognition**]
6) Have people been discussing some sports while talking about the Tokyo Olympics? [**Named Entity Recognition**]
7) If yes, which sports and can they be linked to a knowledge base(s)? [**Named Entity Linking**]

Each of the research questions will utilise a stage(s) from the Social Media Analytics (*SMA*) pipeline. The theory and background regrading these stages is discussed further in the next section (*SMA Pipeline: Background*).

Asking the right questions: With almost *5700* tweets being sent out in a single second on average [9], *Twitter* was in our opinion the best social medium to aid our research efforts; the access to big data makes it easier to interpret whether the proposed questions have a popular opinion. For example, in August 2013 in Japan, *143,199* tweets per second were sent out because twitter users were watching an airing of popular movie *Castle in the Sky* [9]. People were sharing their opinions on the movie, and with so much data *volume*, *velocity* and *variability*, the value of such data cannot be ignored [10]. The implications could involve developing movies or television series using similar scenes, soundtracks, actors, etc.

The research questions proposed provide the opportunity to gain public opinion on the Tokyo Olympics, both good and bad. Our findings can be used to aid businesses, health authorities and even the *IOC* curb their plans for the upcoming events. For example, regarding the 1^{st} and 3^{rd} questions, businesses and health authorities could try to exploit or address some of the topics, and also identify if their current actions are resonating with people. The 5^{th} question will also help illustrate the countries, organisations, personalities, events, or places the public has been discussing the most; this information could benefit television networks, (e.g. Olympic merchandise) businesses and even the athletes.

II. SMA Pipeline: Background

This section goes into depth about the stages within the *SMA* pipeline utilised:

- Data Collection & Preprocessing
- Topic Modelling
- Sentiment Analysis
- Named Entity Recognition (*NER*)
- Named Entity Linking (*NEL*)

A. Data Collection & Preprocessing

Data gathering typically involves choosing a data source such as a social medium/media or human subjects themselves. In our case, *Twitter* is used; the reasons for this are covered in the next section, *Methodology*. Nevertheless, to perform *SMA* fellow researchers have utilised *Twitter* [7], [11], and other media like *Facebook*, *Reddit*, *Instagram* [8]. Efforts have also been made to analyse curated (monolingual or multilingual) datasets [12] and even research paper libraries [13]. The theme common with these methods is usage of open-source data through the use of open application programming interfaces (*API*s) or web crawling [8].

Once the data has been obtained, it needs to be properly formatted to ensure proper functionality of the models for the next four stages. The extent and type of data cleaning to be performed is dependent on the data source and its usage. For example, gathering people's *feelings* or reactions on social media to major events will require decoding the meaning of emojis, such as ':)' or ':(' . On the contrary, identifying hot topics being discussed at major events will rely more on *keywords* like 'shrimp', 'murder', 'Howard'. *Pak et al.* [11] describe a systematic approach to preprocessing data obtained from *Twitter* that we adopt into our *SMA* pipeline; it involves the following:

- Filtering
- Tokenisation
- Stopwords removal
- N-grams construction

Additionally, *Maynard et al.* [12] go into depth about handling emoticons and swear words for sentiment analysis.

B. Topic Modelling

Identifying subjects by using keywords within the dataset is the basis of topic modelling. Normally, it is used to detect topics in *text* documents, this involves grouping both documents that share similar words, and words that occur in similar documents [14]. When applied to social media data, the methodology changes slightly but the end-product is the same.

The most popular technique available is known as Latent Dirichlet Allocation (*LDA*). It is an *unsupervised* (clustering) Machine learning strategy used to extract hidden or *latent* information from text corpus over a *Dirichlet* distribution [15]. It has a number of variants including "*On-line LDA*" [16] and "*labelledLDA*" [12]. Once the allocation has been performed, evaluation needs to be performed; this involves using Normalised Mutual Information (*NMI*), Adjusted Rand Index, F-score, or Entropy [14]. Latent Semantic Analysis (*LSA*) can also be utilised to obtain the "*contextual-usage*" meaning of words from corpus [17]. Its usage significantly improves topic modelling performance, particularly with combating word ambiguity.

Topic modelling is a major stage of the *SMA* pipeline as it helps identify information that may not be explicit in the data. Hidden themes can be determined but it can also be used to aid the direction of the rest of the *SMA* pipeline stages. If a significant theme arises from the results, it can be further analysed to learn even more about the data.

C. Sentiment Analysis

One of the key stages of the *SMA* pipeline, *sentiment analysis* is the tool used to gauge the public mood on specific topics. *Pang et al.* [18] provide a detailed overview of the techniques used to perform the analysis; they identified 3 primary strategies:

1) Lexicon-based
2) Machine learning-based
3) Hybrid-based

Lexicon-based methods involve reference to a prepared database of sentiment terms or keywords i.e. a *lexicon*. When presented with a corpus, sentimental polarity is identified as per the semantic orientation of words or phrases in the predefined lexicon [8].

Machine learning-based approaches include a wide variety of techniques ranging from expensive *supervised* learning with the use of labelled data, to *unsupervised* learning and *deep learning*-based techniques. This involves the use of text classifiers like the *multinomial Naïve Bayes* classifier exhibited by Pak and Paroubek [11]. *Glorot et al.* showcase how a *deep neural network* and a *Denoising Auto-Encoder* can be adapted to extract key sentiment information from corpora [19].

Hybrid-based techniques involve a combination of the two previous approaches. *Modhaddam et al.* [20] demonstrate how the addition of a *sentiment lexicon* to machine-learning based approaches results in an increase in accuracy. They utilise a tuple of values that consist of similarity scores for given adjectives against three predefined adjectives: *excellent*, *mediocre* and *poor*. This sort of approach has become common in current sentiment analysis research, and is adopted as part of our methodology. The main reason for its relevance is that a high accuracy can be obtained with a small training set.

Sentiment analysis is an effective technique to measure public opinion on any subject matter so long as relevant data exists. Social media platforms serve as a *"gold mine"* because of the vast volume of data generated through the opinions and emotions shared by users. Performing large-scale sentiment analysis "ethically" is crucial and when done correctly is beneficial to businesses, advertisers and even the "*average Joe*".

D. Named Entity Recognition

Named Entity Recognition (NER) is a key component in NLP systems for question answering, information retrieval, relation extraction, etc. The Named Entity Recognition approaches can be roughly divided into two types: Flat Named Entity Recognition and Nested Named Entity Recognition.

The majority of flat NER models are based on a sequence labelling approach. Collobert et al. [21] proposed a neural NER model that uses CNNs to encode tokens combined with a CRF layer for the classification. Similarly, Lample et al. [22] uses LSTMs to encode the input and a CRF for the prediction.

For Nested Named Entity Recognition, the named entities can be nested. Ju et al. [23] described a LSTM-CRF model to predict nested named entities. Their algorithm will detect entities iteratively until no further entities can be predicted. Strakova et al. [24] recognize the nested named entity by a sequence-to-sequence model exploring combinations of context-based embeddings such as ELMo, BERT, and Flair.

There is no doubt that in the NER field, Attention Model, Transfer Learning and Semi-Supervised Learning are gradually replacing the traditional CNN-CRF and LSTM-CRF method. Hao et al. [25] presented a semi-supervised framework for transferable NER, which disentangles the domain-invariant latent variables and domain-specific latent variables. He et al. [26] proposed a named entity recognition method that combines knowledge graph embedding with a self-attention mechanism.

E. Named Entity Linking

Due to the fact that entities, such as people, places, and events, can be referred to by many mention strings, and the same mention string may be used to refer to multiple entities. For example, "Georgia" can be a country, a state of USA, even an actress's name "Georgia Tennant". Therefore, linking the mentions to a knowledge base is proposed for disambiguation, which makes it easier for language processing systems to collect and exploit information about entities across documents.

Before Wikipedia, it was nearly impossible to achieve wide-coverage NEL since there was no general purpose, publicly available collection of information about entities. But now, several research communities have proposed a lot of methods to address entity ambiguity problem by named entity linking on Wikipedia. Mihalcea and Csomai [27] used Wikipedia as a word sense disambiguation data set by attempting to reproduce the links between pages. Bunescu and Pasca [28] used Wikipedia in a similar way, but include ner as a preprocessing step and require a link or (nil) for all identified mentions. Yamada et al. [29] proposed a novel method to enhance the performance of the Twitter NER task by using Entity Linking based on Wikipedia.

III. Methodology

This section covers each stage of our implementation of the *SMA* pipeline, referencing some of the concepts discussed in the previous section.

The data used for the work presented in this paper was crawled from the *Twitter* web platform. Efforts where initially invested in using tweepy[1] *Python* library to search for tweets, but the Twitter API is not flexible for obtaining historical data because only tweets between a period of 7 days can be pulled at once using the free API keys [30].

We used an open-source python package, snscrape[2] v0.3.4 to search through Twitter for tweets between October 2019 and March 2021 that included any of the words "#TokyoOlympics", "Tokyo Olympics", "#Tokyo2020" and "#Tokyo2021". As snscrape does not provide the feature to filter tweets by language, our initial search resulted in over *600,000* tweets. Further, the non-English tweets were pruned, and we found that only 52% of the original tweets, i.e. about *317,000* tweets, satisfy our work's scope.

We considered *ethical* standards to ensure that the method used to source data from Twitter is legitimate and in compliance with the Twitter rules and policies. We did an extensive check to guarantee that it is permitted to use the snscrape python library to scrape tweets from Twitter. We also checked the "robots.txt"[3] [31] file on Twitter website to assure that it is permitted to scrape the particular URLs where we extracted the data.

The unit of analysis in this work is a single tweet. Tweets are normally short, and this means it will usually be required for users to include abbreviations, phonetic substitutions, emoticons, emojis and ungrammatical structures that torments text-processing tools [32]. To reach reliable answers to the research questions, we performed some data preprocessing steps. Preprocessing steps included back and forth of processes that warranted a continuous visualisation of the current state of the dataset using wordcloud [4].

1) Punctuations, unnecessary line breaks, contiguous spaces and trailing spaces were removed, and all words were converted to their lowercase equivalent.
2) Usernames, i.e. words that begin with the "@" symbols, e.g. @jack; hashtags, i.e. words that start with the "#" symbol, e.g. #TokyoOlympics; web addresses, e.g. https://twitter.com; Extensible Markup Language (XML) tags, e.g. <title> or </title>; and duplicate tweets where all removed from the text primarily using regular expression [5]
3) Lemmatisation was performed using NLTK [6]. This involves the replacement of grammatical ending with other grammatical end of the normalised word [33] to remove anomalies

[1] https://www.tweepy.org

[2] https://github.com/JustAnotherArchivist/snscrape

[3] https://twitter.com/robots.txt

[4] https://pypi.org/project/wordcloud/

[5] https://docs.python.org/3/library/re.html

[6] https://www.nltk.org/_modules/nltk/stem/wordnet.html

4) Emojis were replaced with their word equivalents using emoji [7] to assist the detection of higher sentiment scores [34]
5) NLTK[8] stopwords corpus was used to remove stopwords from the dataset. Removal of stop words is motivated because they are not measured as keywords in the text mining pipelines [35], and they make the text heavier [36]
6) Abbreviations and slangs were replaced by their full forms using a dictionary of popularly used Twitter abbreviations and slangs.

To get the significant topics discussed by people talking about Tokyo Olympics, we used the Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) for topic modelling. We later used pyLDAvis [9] to extract information from the fitted topic models to aid visualisation and interpretation.

For sentiment analysis, we used three open-source libraries: VADER[10], TextBlob[11] and Text2emotion[12]. VADER is to generate ratios for proportions of text that fall in different categories, including positive, neutral and negative sentiments. The sentiment property in TextBlob returns two values: polarity, subjectivity, where polarity represents the degree of whether a sentence is positive or negative, and subjectivity refers to personal feelings, which indicates whether a sentence is subjective or objective. We further used Text2emotion to extract five different emotions, happy, sad, angry, surprise and fear, from the content. The results are then visualised using Matplotlib's[13] pie chart to enable us to obtain detailed information related to people's feelings towards the Tokyo Olympics between different periods.

To identify the sports that are being discussed, an EntityLinker component was added to the end of spaCy's "en_core_web_trf" pre-trained pipeline to link the recognised entities into a wiki database.

IV. Experimentation: Results and Analysis

This section provides an interpretation of our findings. Visual aids are also referenced within this section as *Figures* and are available in the *Appendices* section of the report.

Firstly, we used LDA and LSA to complete topic construction and to predict the 50 most discussed topics in the corpus, and then used pyLDAvis to visualise the topics to aid interpretation. We observed that both LDA and LSA can classify large corpus into various topics, but the conclusion of the topics from LDA are directly related to the Olympic Games while those from LSA are generic. Similar to a sample in Fig. 12, the result of this pipeline pointed us to the topics that are discussed by people talking about the Tokyo Olympics.

We did sentiment analysis and found that people's feelings towards the Tokyo Olympics changed over time and were largely influenced by different significant events. From October 2019 to March 2020, i.e., the early stage of the COVID-19 pandemic, people held the highest proportion (almost 46%) of positive views compared to the almost 26% of negative opinions about the Tokyo Olympics (see Fig. 1a). A further investigation of the proportion of fear, happiness and sadness shows that 38% of people felt happy about the Olympics, while around 27% and 9% of people expressed fear and sadness, respectively (see Fig. 1b). However, from April 2020 to September 2020, i.e., the period when COVID-19 was spreading worldwide, people's positive sentiments remarkably decreased, and the percentage of happy tweets went down to 8% (see Fig. 2b). Besides, there was an increase in the number of tweets that delivered negative emotions, rising to almost 50% of the corpus (see Fig. 2a). Interestingly, in the last period, from October 2020 to March 2021, the proportion of sad tweets increased by 6%, and the number of happy tweets remained almost the same as the last stage (see Fig. 3). This may have been caused by Japan's announcement that the Tokyo Olympics will take place without overseas spectators.

We searched through the corpus for tweets containing any of the words; "COVID", "corona virus", "covid19", "covid-19", "corona", "pandemic", "roan", "#COVID19", "coronavirus" and "virus". We found that about 15% of the dataset, i.e. almost *45,000* tweets contain at least one of these words. Further, we separated the tweets that contain any of those words because the presence of those words strongly suggests an expression of reaction to the effect of COVID-19 on the Tokyo 2020 Olympics. We then performed sentiment analysis on this subcorpus and found that slightly over 43%, 29% and 27% have negative, positive and neutral sentiments, respectively (see Fig. 4a). To conclude, we were curious to know more than the percentages of positive, negative and neutral contents. We further investigated the portions of contents that suggest happy, sad, angry, surprise and fear emotions. We found that contents revealing fear and anger emotions constitute slightly over 35% of the subcorpus, 31.70% shows happy emotion, 23.20% of the contents shows surprise emotions, and 9.80% shows the sad emotion (see Fig. 4b). Our opinion is that the combined dominance of negative sentiments, i.e. sadness, fear, surprise and anger over the happy sentiments show people's worries and fear of the aftermath and repercussion of holding the Olympic games even amidst the COVID-19 pandemic.

We used spaCy to recognise the named entities in the corpus, counted the words' frequency and plotted graphs for the interesting entities categories. As Fig. 5 indicates, the most discussed person is Yoshiro Mori [15], the last Tokyo Olympic's chief who had to resign in February amid backlash over sexist remarks towards women. It was not a surprise that Japan appeared to be the second most discussed place, following Tokyo, which is the most popular city in Japan (see Fig. 6). As can be seen in Fig. 9, the top 5 discussed nationalities

[7] https://pypi.org/project/demoji
[8] https://www.nltk.org
[9] https://github.com/bmabey/pyLDAvis
[10] https://pypi.org/project/vaderSentiment
[11] https://textblob.readthedocs.io/en/dev/install.html
[12] https://pypi.org/project/text2emotion/
[13] https://matplotlib.org/
[15] https://en.wikipedia.org/wiki/Yoshir%C5%8D_Mori

are Japanese, Indian, Russian, Canadian and Chinese. Interestingly, it is notable from Fig. 7 that the *World Cup* is the most discussed event after excluding topics related to the Olympics. From Fig. 8, we see that the most discussed organisations are two news agencies, Tokyo news and NBC news. It also appears that Airbnb, which is an online marketplace for lodging, is the third most discussed. We suppose that Airbnb is ideal for people who want to travel to Tokyo during the Olympics.

We tried to detect the most discussed sports using spaCy, but there are no present embedded methods in spaCy to achieve this. So we made use of the wiki knowledge base to build a pipeline. This was done by first getting the linked entity of the word ”basketball” and then the superclasses of the entity ”basketball”. One of these superclasses, ”type of sport”, was then used to get its subclasses that are sports names that we used as patterns to recognise the ”SPORT” entities. Figures 10 and 11 show the proportion and named entity linking of the most discussed sports by people discussing the Tokyo Olympics on Twitter respectively.

V. Conclusion

This report went into thorough detail regarding the “Social Media Analytics” (*SMA*) pipeline, and its application on data related to the 2020 Summer Olympic games. The stages of the *SMA* pipeline were reviewed, and an overview behind the theory and practice of these stages was provided. Based on the Olympics games data obtained from the social media platform, *Twitter*, we defined our implementation of the *SMA* pipeline. The techniques covered in the background were accordingly adopted and utilised to extract key information from the social media data. Following this, an interpretation of the results obtained was provided with additional visual aids to strengthen our deductions. “COVID-19” was determined to significantly influence the emotions of the public. Utilising topic modelling and *NER*, latent information regarding the Olympics was uncovered including the most talked about countries and people too. *SMA* is a very promising research area and our findings further evidence this. This paper has demonstrated that a rigorous analysis of social media data can be utilised to conduct a survey of public opinion on specific matters, setting the stage to learn even more about data.

References

[1] M. B. Duignan, “Leveraging tokyo 2020 to re-image japan and the olympic city, post-fukushima,” *Journal of Destination Marketing & Management*, vol. 19, p. 100486, 2021.

[2] J. Millán-Oñate, A. J. Rodriguez-Morales, G. Camacho-Moreno, H. Mendoza-Ramírez, I. A. Rodríguez-Sabogal, and C. Álvarez-Moreno, “A new emerging zoonotic virus of concern: the 2019 novel coronavirus (covid-19),” *Infectio*, vol. 24, no. 3, pp. 187–192, 2020.

[3] V. Oblinger-Peters and B. Krenn, ““time for recovery” or “utter uncertainty”? the postponement of the tokyo 2020 olympic games through the eyes of olympic athletes and coaches. a qualitative study,” *Frontiers in Psychology*, vol. 11, p. 3619, 2020.

[4] I. Cerullo, “Impact and meaning of the postponement of tokyo 2020 from an athlete perspective,” *Olimpianos-Journal of Olympic Studies*, vol. 4, pp. 28–36, 2020.

[5] Z. Shervani, I. Khan, U. Y. Qazi *et al.*, “Sars-cov-2 delayed tokyo 2020 olympics: Very recent advances in covid-19 detection, treatment, and vaccine development useful conducting the games in 2021,” *Advances in Infectious Diseases*, vol. 10, no. 03, p. 56, 2020.

[6] S. Aral, C. Dellarocas, and D. Godes, “Introduction to the special issue—social media and business transformation: a framework for research,” *Information Systems Research*, vol. 24, no. 1, pp. 3–13, 2013.

[7] P. Brooker, J. Barnett, and T. Cribbin, “Doing social media analytics,” *Big Data & Society*, vol. 3, no. 2, p. 2053951716658060, 2016.

[8] B. Jeong, J. Yoon, and J.-M. Lee, “Social media mining for product planning: A product opportunity mining approach based on topic modeling and sentiment analysis,” *International Journal of Information Management*, vol. 48, pp. 280–290, 2019.

[9] R. Krikorian, “New tweets per second record, and how,” *Twitter Official Blog*, vol. 221, 2013.

[10] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton, “Big data: the management revolution,” *Harvard business review*, vol. 90, no. 10, pp. 60–68, 2012.

[11] A. Pak and P. Paroubek, “Twitter based system: Using twitter for disambiguating sentiment ambiguous adjectives,” in *Proceedings of the 5th International Workshop on Semantic Evaluation*, 2010, pp. 436–439.

[12] D. Maynard, K. Bontcheva, and D. Rout, “Challenges in developing opinion mining tools for social media,” *Proceedings of the@ NLP can u tag# usergeneratedcontent*, pp. 15–22, 2012.

[13] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, “Social media analytics–challenges in topic discovery, data collection, and data preparation,” *International journal of information management*, vol. 39, pp. 156–168, 2018.

[14] S. A. Curiskis, B. Drake, T. R. Osborn, and P. J. Kennedy, “An evaluation of document clustering and topic modelling in two online social networks: Twitter and reddit,” *Information Processing & Management*, vol. 57, no. 2, p. 102034, 2020.

[15] D. A. Ostrowski, “Using latent dirichlet allocation for topic modelling in twitter,” in *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*. IEEE, 2015, pp. 493–497.

[16] J. H. Lau, N. Collier, and T. Baldwin, “On-line trend analysis with topic models:# twitter trends detection topic model online,” in *Proceedings of COLING 2012*, 2012, pp. 1519–1534.

[17] T. K. Landauer, P. W. Foltz, and D. Laham, “An introduction to latent semantic analysis,” *Discourse processes*, vol. 25, no. 2-3, pp. 259–284, 1998.

[18] B. Pang and L. Lee, “Opinion mining and sentiment analysis,” *Found. Trends Inf. Retr.*, vol. 2, no. 1–2, p. 1–135, Jan. 2008. [Online]. Available: https://doi.org/10.1561/1500000011

[19] X. Glorot, A. Bordes, and Y. Bengio, “Domain adaptation for large-scale sentiment classification: A deep learning approach,” in *ICML*, 2011.

[20] S. Moghaddam and F. Popowich, “Opinion polarity identification through adjectives,” *arXiv preprint arXiv:1011.4623*, 2010.

[21] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, “Natural language processing (almost) from scratch,” *Journal of machine learning research*, vol. 12, no. ARTICLE, pp. 2493–2537, 2011.

[22] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, “Neural architectures for named entity recognition,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 260–270. [Online]. Available: https://www.aclweb.org/anthology/N16-1030

[23] M. Ju, M. Miwa, and S. Ananiadou, “A neural layered model for nested named entity recognition,” in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 1446–1459.

[24] J. Straková, M. Straka, and J. Hajič, “Neural architectures for nested ner through linearization,” *arXiv preprint arXiv:1908.06926*, 2019.

[25] Z. Hao, D. Lv, Z. Li, R. Cai, W. Wen, and B. Xu, “Semi-supervised disentangled framework for transferable named entity recognition,” *Neural Networks*, vol. 135, pp. 127–138, 2021.

[26] S. He, D. Sun, and Z. Wang, “Named entity recognition for chinese marine text with knowledge-based self-attention,” *Multimedia Tools and Applications*, pp. 1–15.

[27] R. Mihalcea and A. Csomai, “Wikify! linking documents to encyclopedic knowledge,” in *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, ser. CIKM ’07. New York, NY, USA: Association for Computing Machinery, 2007, p. 233–242. [Online]. Available: https://doi.org/10.1145/1321440.1321475

[28] R. Bunescu and M. Pasca, “Using encyclopedic knowledge for named entity disambiguation,” 2006.

[29] I. Yamada, H. Takeda, and Y. Takefuji, “Enhancing named entity recognition in twitter messages using entity linking,” in *Proceedings of the Workshop on Noisy User-generated Text*, 2015, pp. 136–140.

[30] W. Ahmed, P. A. Bath, and G. Demartini, “Using twitter as a data source: An overview of ethical, legal, and methodological challenges,” *The Ethics of Online Research*, 2017.

[31] C. Yang and H.-J. Liao, “Using the robots. txt and robots meta tags to implement online copyright and a related amendment,” *Library hi tech*, 2010.

[32] R. Sproat, A. W. Black, S. Chen, S. Kumar, M. Ostendorf, and C. Richards, “Normalization of non-standard words,” *Computer speech & language*, vol. 15, no. 3, pp. 287–333, 2001.

[33] D. Mladenic, “Learning word normalization using word suffix and context from unlabeled data,” in *ICML*, 2002.

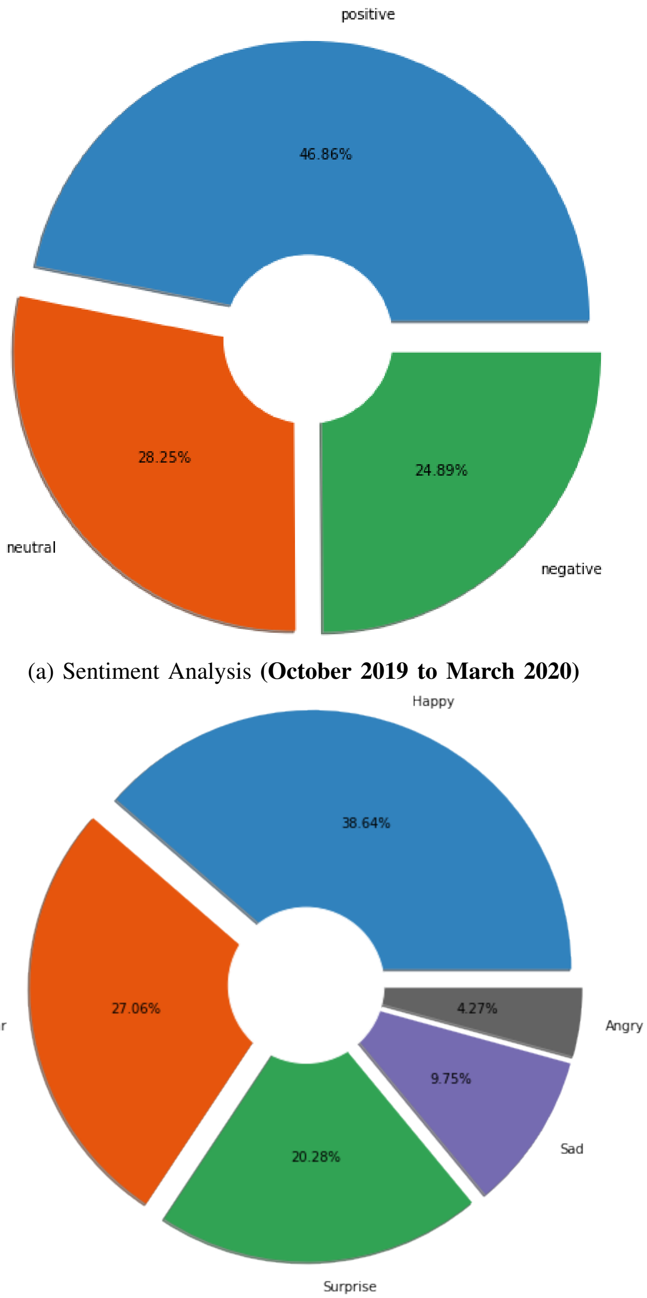
[34] M. Shiha and S. Ayvaz, “The effects of emoji in sentiment analysis,” *Int. J. Comput. Electr. Eng.(IJCEE.)*, vol. 9, no. 1, pp. 360–369, 2017.

[35] M. F. Porter, “An algorithm for suffix stripping,” *Program*, 1980.

[36] S. Vijayarani, M. J. Ilamathi, M. Nithya *et al.*, “Preprocessing techniques for text mining-an overview,” *International Journal of Computer Science & Communication Networks*, vol. 5, no. 1, pp. 7–16, 2015.

VI. Appendices

A. Appendix A

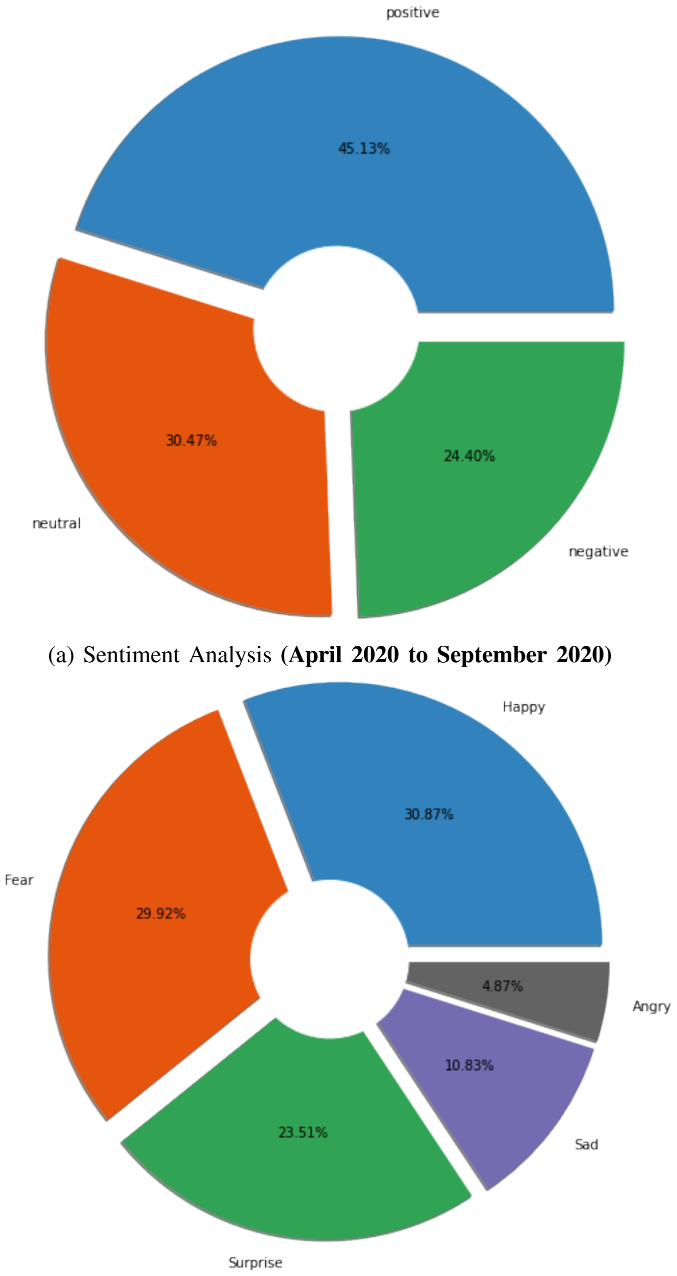


(a) Sentiment Analysis **(October 2019 to March 2020)**

(b) Emotion Analysis **(October 2019 to March 2020)**

Fig. 1: Emotion and sentiment analysis result on the tweets for the first period **(October 2019 to March 2020)**

B. *Appendix B*

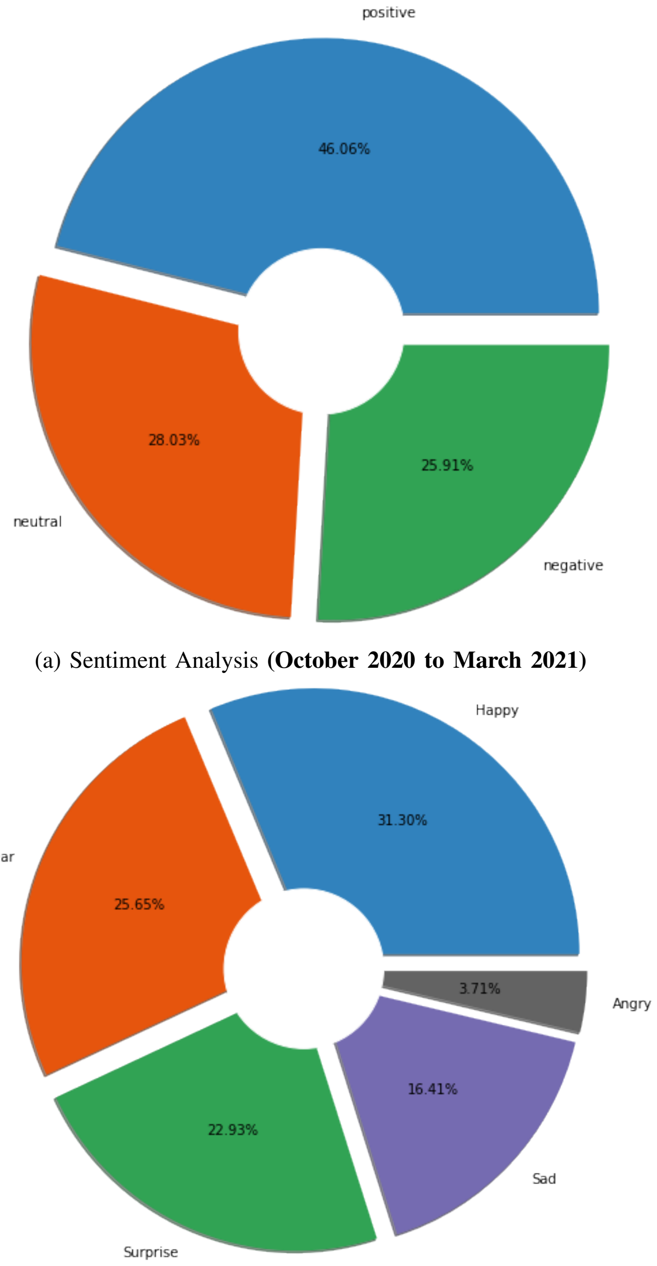


(a) Sentiment Analysis **(April 2020 to September 2020)**

(b) Emotion Analysis **(April 2020 to September 2020)**

Fig. 2: Emotion and sentiment analysis result on the tweets for the second period **(April 2020 to September 2020)**

C. *Appendix C*

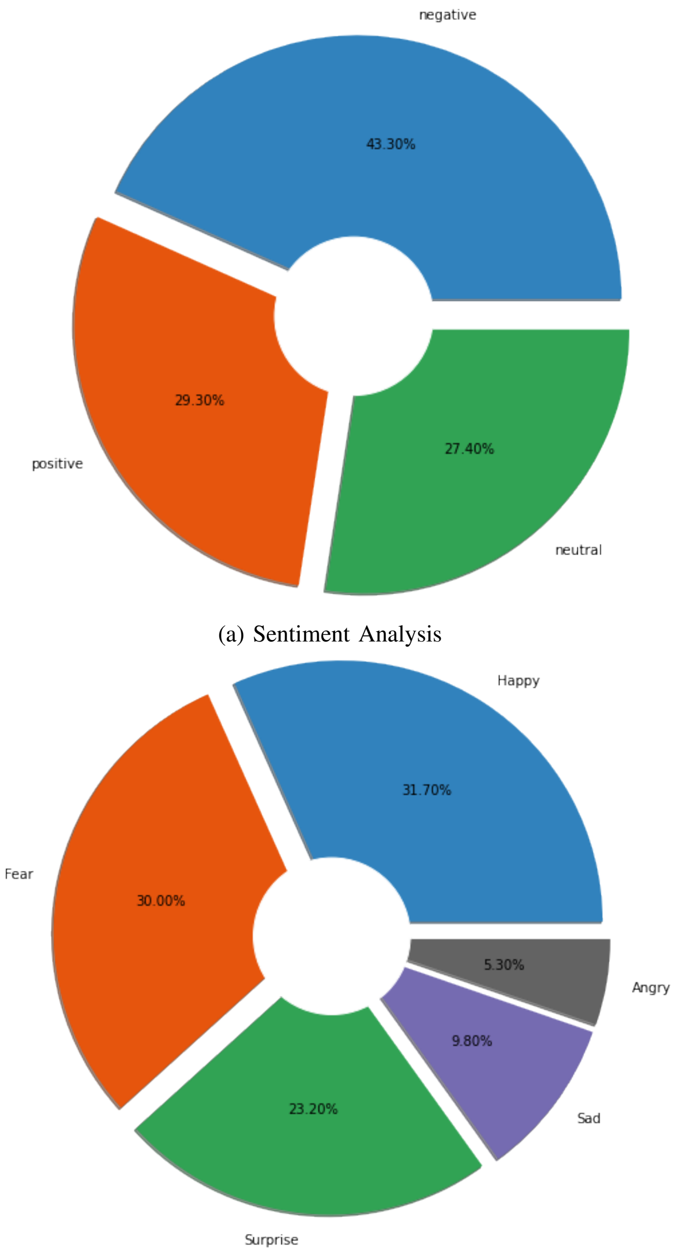


(a) Sentiment Analysis **(October 2020 to March 2021)**

(b) Emotion Analysis **(October 2020 to March 2021)**

Fig. 3: Emotion and sentiment analysis result on the tweets for the last period **(October 2020 to March 2021)**

D. Appendix D



(a) Sentiment Analysis

(b) Emotion Analysis

Fig. 4: Proportion of emotions in tweets containing texts that strongly suggests the **pandemic** is being discussed

E. Appendix E

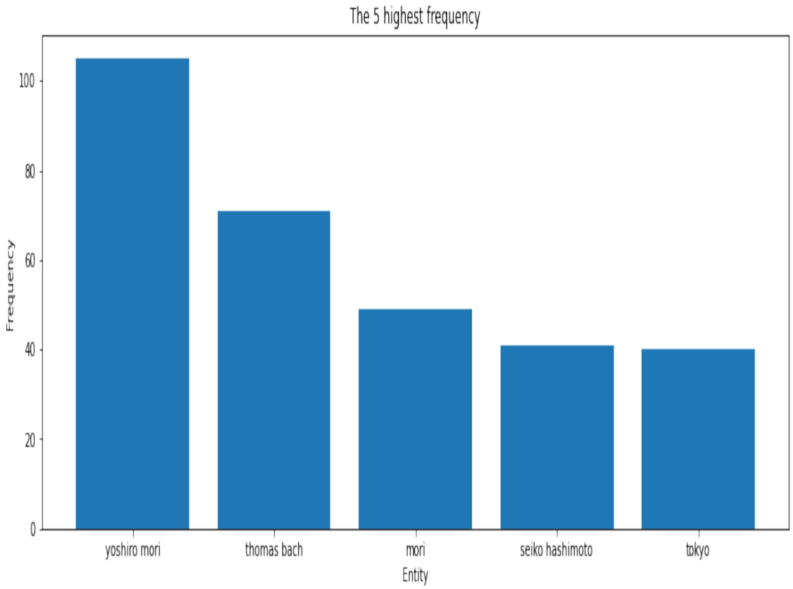


Fig. 5: Proportion of the most discussed **persons**

F. Appendix F

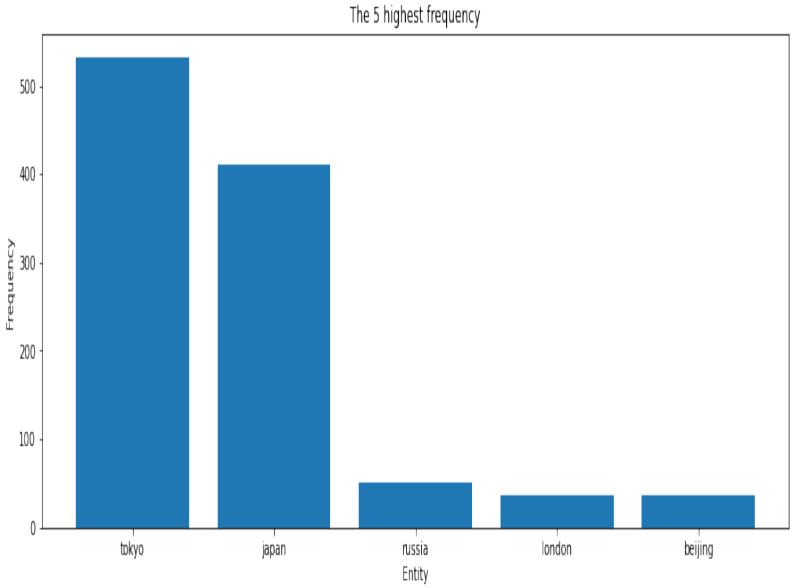


Fig. 6: Proportion of the most discussed **geopolitical entities**

G. Appendix G

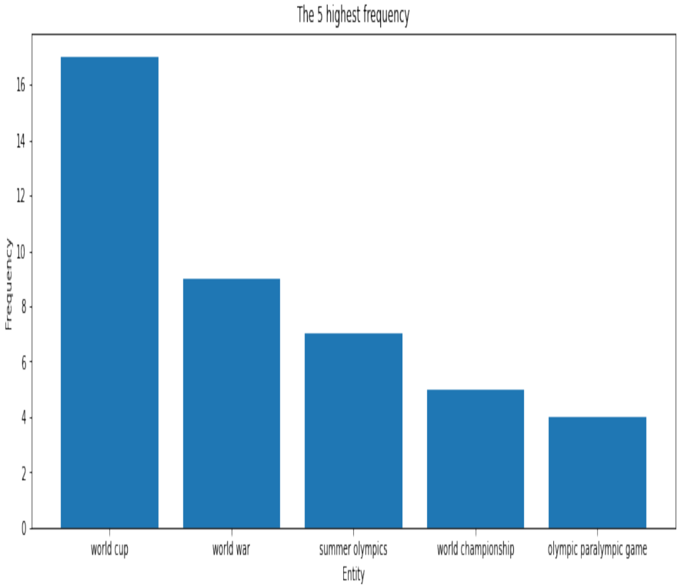


Fig. 7: Proportion of the most discussed **events**

H. Appendix H

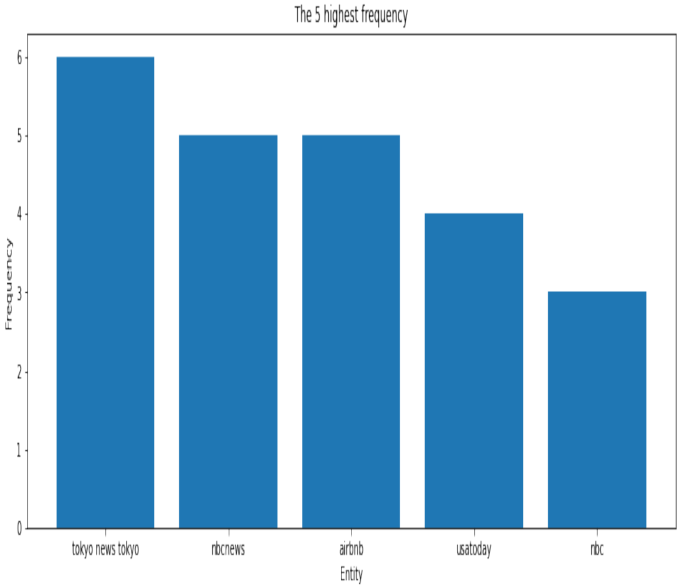


Fig. 8: Proportion of the most discussed **organisations**

I. Appendix I

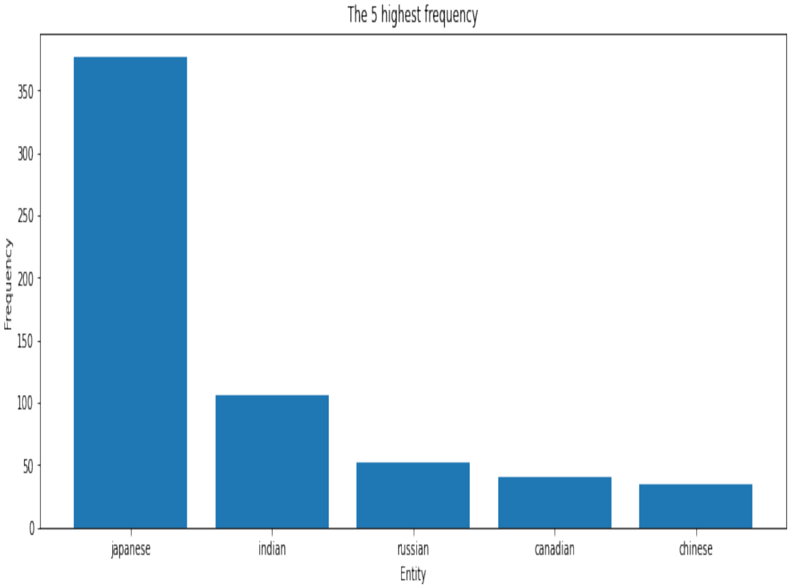


Fig. 9: Proportion of the most discussed **nationalities, religious or political groups**

J. Appendix J

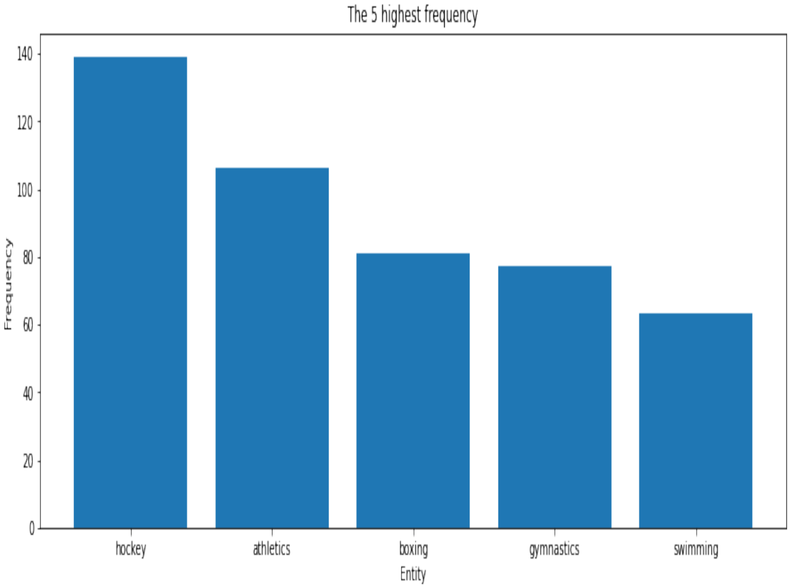
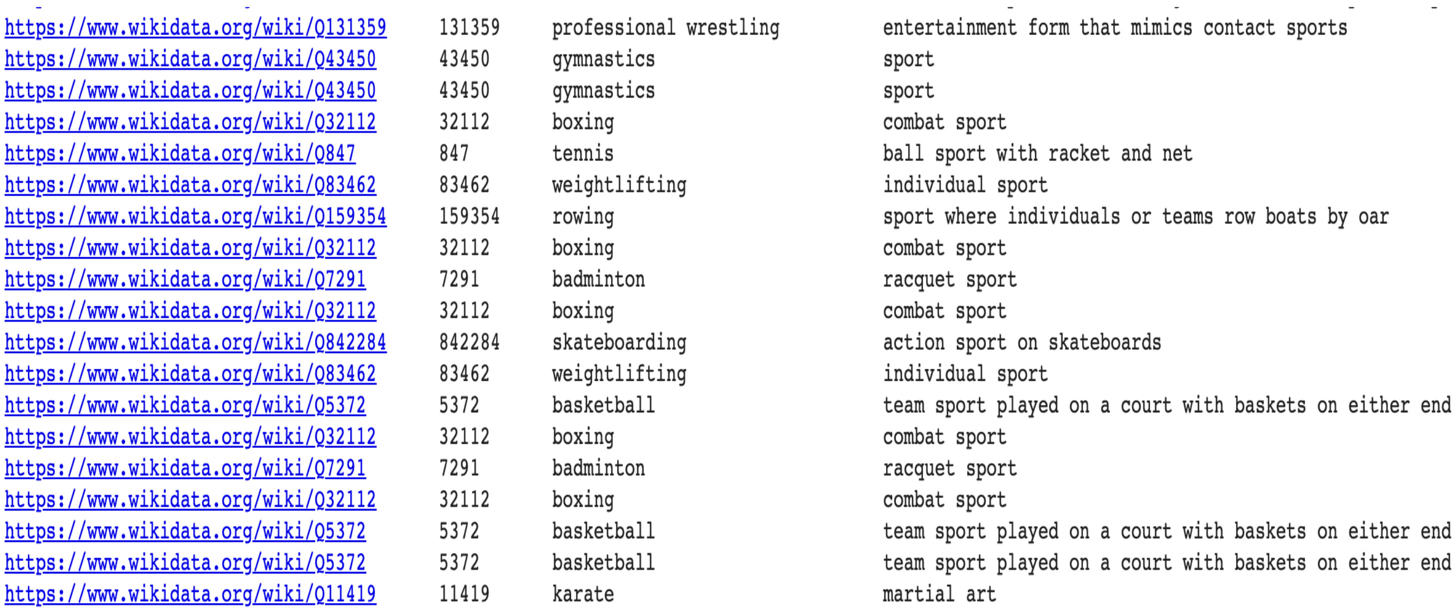


Fig. 10: Proportion of the most discussed **sports**

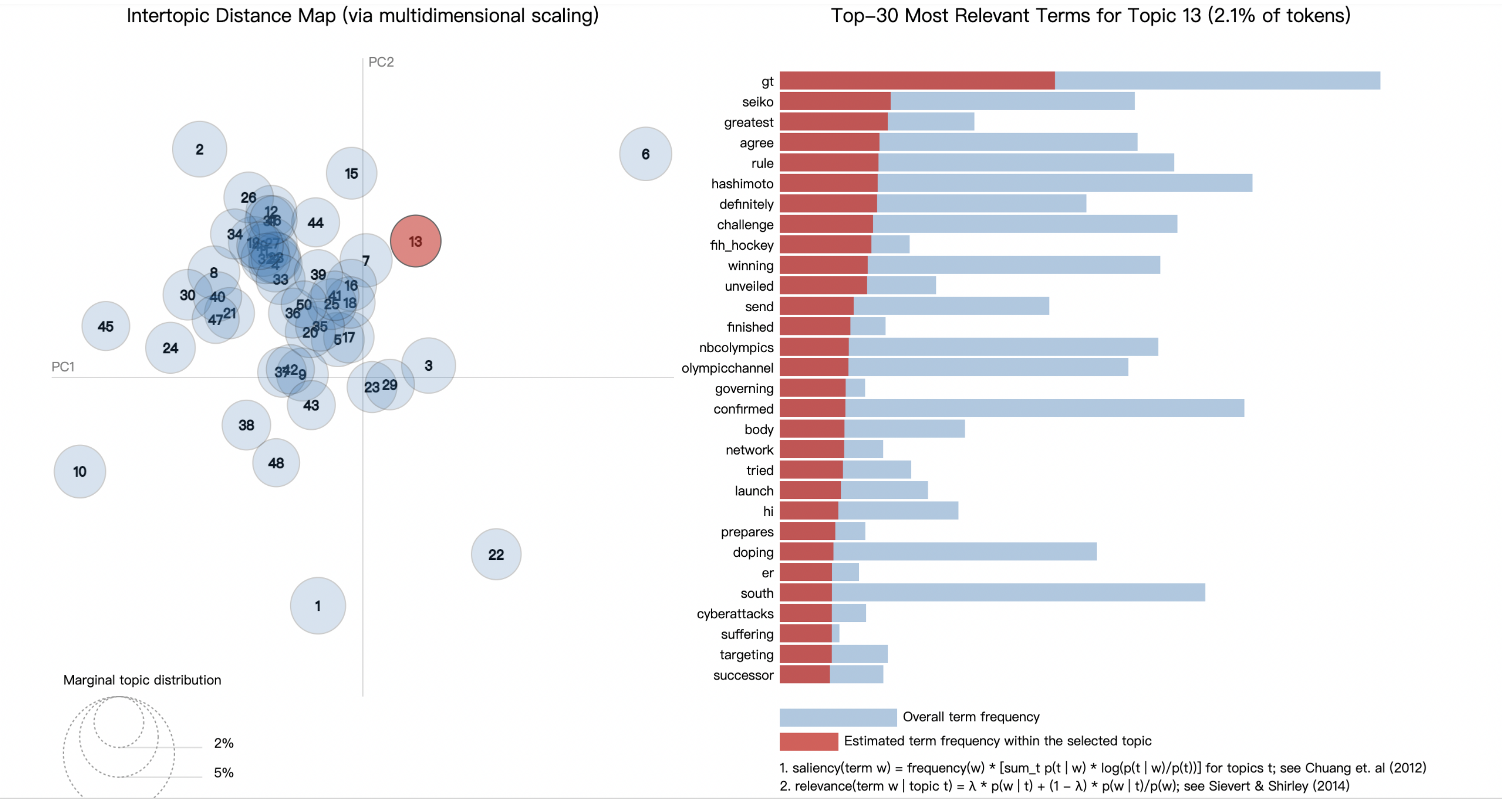
K. Appendix K



https://www.wikidata.org/wiki/Q131359	131359	professional wrestling	entertainment form that mimics contact sports
https://www.wikidata.org/wiki/Q43450	43450	gymnastics	sport
https://www.wikidata.org/wiki/Q43450	43450	gymnastics	sport
https://www.wikidata.org/wiki/Q32112	32112	boxing	combat sport
https://www.wikidata.org/wiki/Q847	847	tennis	ball sport with racket and net
https://www.wikidata.org/wiki/Q83462	83462	weightlifting	individual sport
https://www.wikidata.org/wiki/Q159354	159354	rowing	sport where individuals or teams row boats by oar
https://www.wikidata.org/wiki/Q32112	32112	boxing	combat sport
https://www.wikidata.org/wiki/Q7291	7291	badminton	racquet sport
https://www.wikidata.org/wiki/Q32112	32112	boxing	combat sport
https://www.wikidata.org/wiki/Q842284	842284	skateboarding	action sport on skateboards
https://www.wikidata.org/wiki/Q83462	83462	weightlifting	individual sport
https://www.wikidata.org/wiki/Q5372	5372	basketball	team sport played on a court with baskets on either end
https://www.wikidata.org/wiki/Q32112	32112	boxing	combat sport
https://www.wikidata.org/wiki/Q7291	7291	badminton	racquet sport
https://www.wikidata.org/wiki/Q32112	32112	boxing	combat sport
https://www.wikidata.org/wiki/Q5372	5372	basketball	team sport played on a court with baskets on either end
https://www.wikidata.org/wiki/Q5372	5372	basketball	team sport played on a court with baskets on either end
https://www.wikidata.org/wiki/Q11419	11419	karate	martial art

Fig. 11: Named entity linking for some of the discussed sports

L. Appendix L



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 – λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Fig. 12: Topic Modelling: Most relevant terms for topic #13