Deriving Disinformation Insights from Geolocalized Twitter Callouts

David Tuxworth
Crime & Security Research Institute
School of Computer Science and Informatics
Cardiff University
Cardiff, UK
tuxworthdt@cardiff.ac.uk

Dimosthenis Antypas
Crime & Security Research Institute
School of Computer Science and Informatics
Cardiff University
Cardiff, UK
antypasd@cardiff.ac.uk

Luis Espinosa-Anke
Crime & Security Research Institute
School of Computer Science and Informatics
Cardiff University
Cardiff, UK
espinosa-ankel@cardiff.ac.uk

Jose Camacho-Collados
Crime & Security Research Institute
School of Computer Science and Informatics
Cardiff University
Cardiff, UK
camachocolladosj@cardiff.ac.uk

Alun Preece
Crime & Security Research Institute
School of Computer Science and Informatics
Cardiff University
Cardiff, UK
preecead@cardiff.ac.uk

David Rogers
Crime & Security Research Institute
School of Computer Science and Informatics
Cardiff University
Cardiff, UK
rogersdm1@cardiff.ac.uk

ABSTRACT

This paper demonstrates a two-stage method for deriving insights from social media data relating to disinformation by applying a combination of geospatial classification and embedding-based language modelling across multiple languages. In particular, the analysis in centered on Twitter and disinformation for three European languages: English, French and Spanish. Firstly, Twitter data is classified into European and non-European sets using BERT. Secondly, Word2vec is applied to the classified texts resulting in Eurocentric, non-Eurocentric and global representations of the data for the three target languages. This comparative analysis demonstrates not only the efficacy of the classification method but also highlights geographic, temporal and linguistic differences in the disinformation-related media. Thus, the contributions of the work are threefold: (i) a novel language-independent transformer-based geolocation method; (ii) an analytical approach that exploits lexical specificity and word embeddings to interrogate user-generated content; and (iii) a dataset of 36 million disinformation related tweets in English, French and Spanish.

1 INTRODUCTION

Social media provides a rich stream of user-generated data that can be utilised in many ways. This paper employs a two-stage method to use this resource in order to derive insights into disinformation. The scale, immediacy and popularity of social media render it an ideal platform for the dissemination of ideas. While the many platforms available are used for legitimate communication, it is also used by modern propagandists to wilfully spread false information, i.e., disinformation. The inadvertent sharing of false information, i.e., misinformation, is widespread and while not necessarily malicious in intent, can be hugely damaging. Understanding the content targeted at as well as generated by users of social media is paramount in tackling these phenomena. Computational methods are required not only to analyze but to keep pace with the volume of data generated by both legitimate and illegitimate users of social media. A further challenge is considering the language, culture and context of the messaging. These elements are considered in this paper.

The motivation for this study is practical, embedded in ongoing work to detect, track and understand disinformation operations in a variety of geopolitical contexts. To this end, Twitter data relating to misinformation, disinformation and related terms including propaganda and 'fake news' have been continuously collected since 2019 in multiple languages including English, French and Spanish, which are the languages of focus in this study. The intuition behind the collection method is that Twitter users often 'call out' misinformation and disinformation (following the definitions in [28]) through tagging or quoting media they find questionable. Of course, this does not mean that the media is actually misinformation or disinformation; often it is simply content that the users find objectionable. Nevertheless, collecting data with those terms (translated across the set of target languages) provides a superset of material for analysis. Given the global nature of English, French and Spanish, it becomes necessary to distinguish regional narratives, particularly the Americas versus Europe, from global ones. In turn, examining the use of language around specific query terms such as 'immigrant'/'immigré'/'inmigrante' can help derive insights into mis/disinformation narratives relating to those terms. How the use of language evolves over time is also potentially revealing.

To achieve this, the paper describes a two-stage method by which (1) user-generated data from Twitter is classified into European and non-European subsets in three languages: English, French and Spanish, and (2) embedding-based language models are built for each of the subsets, further subdivided into two periods of time. The choice of languages and time periods are illustrative; the method is completely general. English, French and Spanish were selected as a subset of languages for which data had been collected because all three are 'global' languages relevant in the context of America and Europe. Time periods in 2019 and 2020 were selected because the

WIT2021, August 14–18, 2021, Singapore
.

former covered a period of significant political activity in Europe—the 9th European Parliament Elections, held during the time when the United Kingdom was in the process of leaving the European Union—and the latter covered the run-up to the 59th US Presidential Election; therefore, these two periods could be expected to provide distinctive regional narratives in each case. Moreover, the onset of the global Coronavirus pandemic in early 2020 would likely further differentiate narratives between the two periods, though with potentially less regional difference.

The main contributions of this work are (1) a novel transformer-based geolocation method that performs in multiple languages; and (2) an analytical method that uses lexical specificity and word embeddings to interrogate multilingual user-generated content with respect to mis/disinformation narratives. In addition, a dataset[1] of 36 million disinformation related tweets in English, French and Spanish is made available to researchers.

The paper is structured as follows: Section 2 summarises related work; Section 3 provides details of the multilingual disinformation-related dataset; 4 presents the classification method and performance results; 5 describes the analytical method using lexical specificity and work embeddings; finally, 6 concludes the paper and highlights future work.

2 RELATED WORK

2.1 Twitter Geolocalization

Previous research [1] shows that geotagging literature exists in three categories: network, text and hybrid methods. A user's connections on social media are strong indicators of an individual's location [16] and so it follows that network-based approaches have been highly successful in geolocalizing user locations. Work by [5] approaches the problem by inferring an unknown user's location through their friend's locations via a mention network. This technique is applied at scale in a distributed system enabling a predicted geolocation of millions of users. However, Huang and Chen [17] have shown that exclusively network-based methods cannot geolocate all users, particularly those that do not form connections meaning there is no network structure available.

The problem of geolocalizing non-geotagged tweets has been approached at varying levels of granularity including at the level of city neighborhoods by comparing the content of tweets to known geolocated examples [25]. In this case the geographic regions, European or non-European, are far broader and are more comparable to country-level geolocation which has been shown to be a less challenging problem than city-level geolocation [14].

A hybrid approach, combining both text and network features is recommended by [17] and [1]. This is not possible in this case as the dataset excludes the attributes required to apply a network-based method and the tweet text is filtered by keywords resulting in the choice of employing metadata in the classification stage. It should be noted that the location and description are user-defined and are thus susceptible to data integrity issues whether by omission or using text which is not relevant or inaccurate. Despite this noise, experiments by [12] show that user-supplied locations contain valuable information and classifiers using the location field outperform purely text-based methods when predicting city-level location. In this work both user location and user description are leveraged as text features to feed into a machine learning classifier, using Twitter geolocalized tweets as seeds.

[1] https://github.com/tuxworth/disinformation-insight-twitter

2.2 Deriving Insights from Twitter Data

Concerning the technical aspects relevant to this paper, this subsection focuses on the well-known word embeddings techniques and their applications to content analysis. The Word2vec [23] toolkit, in its two variants CBOW and SkipGram, is one of the best known techniques for learning word embeddings. These dense vector representations have been leveraged extensively, for example, as input representations in neural network architectures for NLP tasks [10], e.g., detecting 'fake news' and phenomena related to the setting of this work [29]. In a recent study identifying online propaganda [18], Word2vec embeddings were found to outperform a multilingual version of BERT in Urdu [7], which the authors ascribe to the limited vocabulary of Urdu in the model. In another study, Word2vec has been leveraged as a feature in the detection of fake news where researchers found that it performs well in comparison to other textual features across multiple datasets and languages [9]. Using ensemble methods to detect fake news, [15] use Word2vec as an embedding layer in a LSTM architecture.

In this paper, the learned embeddings are used to perform comparative analyses between classified sets of text rather than as input for a downstream task. The primary advantage of Word2vec is the ability to learn semantic relations between words via unsupervised machine learning. Word embedding models can be used to learn analogies (comparison between two elements based on limited shared characteristics). In fact, using word vector analogies as a proxy for understanding behaviours in online communities has been the focus, for example, in [19], who used Twitch data to learn word and emoji embeddings which they then use to study Twitch-specific language, or in [8], who studied emoji analogies in Twitter-specific embeddings. Finally, beyond analogies, Twitter embeddings have also been at the center of studies on gender and race [2], as well as detecting semantic shift during the COVID-19 pandemic [11].

3 TWITTER DATA

The data is collected via the Twitter API from two time periods: 2019-04-17 to 2019-06-30 and 2020-04-17 to 2020-6-30 (inclusive of both start and end dates). The 2019 range is selected as it covers the period surrounding the 2019 European Parliament elections that started the 23rd May. The 2020 range is selected to facilitate a comparative analysis between two years. 97 terms across the three languages were selected by subject-matter experts as being indicative of the concept of disinformation including: 'misinformation', 'fake news', 'propaganda', and 'lies'. These terms were used to collect the dataset. Three European languages are analyzed: English, Spanish and French selected by the 'lang' attribute present within the tweet JSON. Figure 1 shows the proportion of tweets per language for the two years. The total number of tweets is 87,894,019 by 14,803,949 unique users.

294,877 tweets contain geolocation metadata which is 0.34% of the total. To split the data into European and non-European

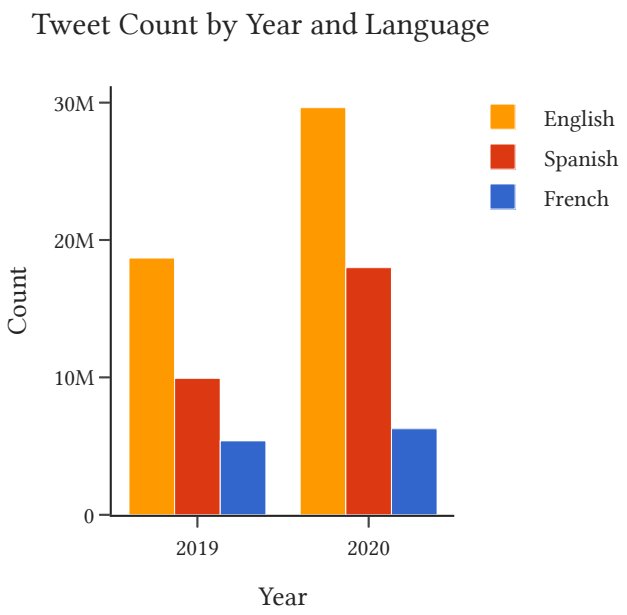

Figure 1: Number of tweets by year and language.

tweets a classifier is trained using the samples that have geolocation data. The classifier is then applied to the remaining tweets that do not contain geolocation data. The class labels are derived from the country code. Tweets with geolocation metadata are labelled European if the country code matches one of those shown in Table 1 and non-European otherwise.

Table 1: ISO 3166 country codes used to select the training data for the European class.

ISO 3166 Country Code
AD, AL, AM, AT, AX, AZ, BA, BE, BG, BY, CH, CY, CZ, DE, DK, EE, ES, FI, FO, FR, GB, GE, GG, GI, GR, HR, HU, IE, IM, IS, IT, JE, KZ, LI, LT, LU, LV, MC, MD, ME, MK, MT, NL, NO, PL, PT, RO, RS, RU, SE, SI, SK, SM, TR, UA, VA

4 GEOLOCALIZATION CLASSIFICATION

As geolocation data is only available for 0.34% of tweets, a method was developed to classify the data into geographic region. This section describes the methodology to attain location information for all tweets in the dataset.

4.1 Experimental Setting

Training and testing data. The subset of tweets which contain geolocation data from the full dataset are used to create a training corpus. Table 2 shows the number of labelled tweets used for the geolocalization classification evaluation (all of them were subsequently used as training data to label the rest of the Twitter corpus). The user location and user description are used as features. For evaluation purposes a 80/10/10 (train/validation/test) stratified split is used for each language dataset.

Table 2: Distribution of tweets used for training the BERT classifiers.

2019			2020		
English	Spanish	French	English	Spanish	French
60,430	49,250	21,816	74,206	66,820	22,355

Preprocessing. A simple pre-processing step is applied to both the user description and user location where punctuation is removed and words (based on letters from the Unicode Basic Latin and Latin-1 Supplement) are extracted. User locations such as 'New York' are concatenated to one term 'new_york'.

Text classification. Following this, a binary classifier is trained for each language using the user description and the user location as features and a Boolean label of 'European' derived from the country code. Initially, a Naive Bayes classifier is used as a baseline model based on the implementation provided from scikit-learn [26]. Then, experiments are carried out with BERT-like models adapted for text classification. In total six models are trained and tested, one for each (language, year) combination.

Pre-trained language models. The BERT-base model [6] is used for the English language, while for Spanish and French BETO [3] and FlauBERT [21] are applied respectively. All models trained are based on the implementations of the uncased versions provided by Hugging Face [31]. Finally, we also experiment with a multilingual BERT model (mBERT).

BERT Optimization. All the BERT models were trained using the same process. Adam optimizer [22] and a linear scheduler with warmup is utilized. We warm up linearly for 500 steps with a learning rate of 5e-5, while a batch size n=34 is used. The models are trained up to 20 epochs, with a checkpoint in every epoch, while an early-stop callback stops the training process after 3 epochs without a performance increase of at least 0.01. We select the best model out of all the checkpoints based on their performance on the dev set.

4.2 Results

As Table 3 shows, the performance of the yearly BERT models is satisfactory for the task at hand with all the models achieving more than 85% accuracy. For both 2019 and 2020 the English model appears to perform better (92% F1-score) while the French model produces the 'worst' results with 87% and 86% F1-score. The difference in the performance could be justified by the smaller training datasets that were available for the Spanish and French languages (Table 2).

Cross-temporal analysis. An effort was made to train and use BERT models only using the 2019 data. The classification metrics when tested on the 2020 data (Table 3: *Bert 2019/2020*), indicate that even though for the Spanish and French datasets the model's performance is on par (same F1 score for Spanish) or even slightly better for French with the models trained on each year, the performance on the English dataset drops (from 92% to 91% F1 score). This shows

Table 3: Classification results for the 2019 and 2020 datasets for each language model. Evaluation metrics: accuracy and macro-averaged precision, recall and F1. mBERT* model is trained on the whole corpus including the three languages. Naive baseline refers to a system where every tweet entry is classified as European

			English				Spanish				French			
Trained	Tested	Classifier	Prec	Rec	Acc	F1	Prec	Rec	Acc	F1	Prec	Rec	Acc	F1
2019	2019	BERT	**0.94**	**0.89**	**0.95**	**0.92**	**0.92**	**0.85**	**0.94**	**0.88**	0.89	0.86	0.9	0.87
		mBERT	0.93	**0.89**	**0.95**	0.91	0.91	0.84	0.93	0.87	**0.91**	**0.87**	**0.91**	**0.89**
		mBERT*	**0.94**	0.86	0.94	0.89	0.51	0.51	0.74	0.51	0.48	0.49	0.64	0.47
		Naive Bayes	0.89	0.81	0.92	0.84	0.88	0.81	0.92	0.84	0.86	0.81	0.87	0.83
2020	2020	BERT	**0.95**	**0.88**	**0.96**	**0.92**	**0.94**	0.84	0.94	0.88	**0.91**	**0.84**	0.89	**0.86**
		mBERT	**0.95**	0.86	0.95	0.9	**0.94**	**0.85**	**0.95**	**0.89**	0.9	0.83	**0.9**	**0.86**
		mBERT*	0.94	0.87	0.95	0.9	0.5	0.5	0.74	0.5	0.49	0.5	0.63	0.47
		Naive Bayes	0.9	0.82	0.92	0.85	0.9	0.82	0.93	0.85	0.86	0.83	0.87	0.84
2019	2020	BERT	0.94	0.89	0.95	0.91	0.92	0.85	0.94	0.88	0.91	0.87	0.91	0.89
		Naive Baseline	0.25	0.5	0.5	0.33	0.25	0.5	0.5	0.33	0.25	0.5	0.5	0.33

that BERT classifiers based on user descriptions are robust even for different periods from where it was trained, which can be relevant for practical settings.

Multilingual BERT. A multilingual BERT model (mBERT) is trained and tested using the combined language datasets for 2019 and for 2020. Unfortunately, training on all languages did not lead to improvements and indeed the results were inferior (see Table 3: *mBERT**). However, the same multilingual model is competitive for all languages when trained on individual language datasets separately. In this case there is an improved performance on the French dataset for 2019 (87% to 89% F1 score) and on the Spanish dataset for 2020 (88% to 89% F1 score) when compared to individual models.

Most of the models trained displayed similar performances when tested. It is possible that by using a different multilingual implementation, or further fine-tuning the existing multilingual model, better results could be achieved compared with using monolingual models across all languages. At the same time, it has been observed in related research [33][32] that for high resources languages, like the ones investigated, mBERT can perform worse than monolingual BERT models depending the task. As the main objective was inferring the location of unseen tweets it was decided to use different models for each language for each year studied. The monolingual BERT model indeed achieved the best results for the largest part of our corpus (English tweets subset). The selected monolingual BERT classifiers are then applied to the rest of the data to create the European and non-European sets. This enables us to analyze the Twitter corpus collected as described in Section 3, with all tweets tagged with location information.

5 ANALYSIS

To enable a balanced comparison between languages, the classified tweet texts are filtered to include only those that match a subset of terms originally used to collect the data. The terms, shown in Table 4, revolve around disinformation, propaganda and themes of influence. Figure 2 shows the classified tweets after applying this step. The total number of tweets is 36,655,061.

Table 4: Terms used to filter tweets for training the embeddings by each language.

English	Spanish	French
active measures	medidas activas	mesures actives
conspiracy	conspiración	complot
deceive	engañar	tromper
deep state	estado profundo	état profond
disinformation	desinformación	désinformation
fabrication	invención	invention
fake news	noticias falsas	fausse nouvelle
influence	influencia	influence
interference	interferencia	ingérence
manipulate	manipular	manipuler
misinformation	desinformación	désinformation
propaganda	propaganda	propagande
subversion	subversión	subversion

The following section describes analyses to derive insights from this geolocalized corpus of tweets, by means of lexical specificity (Section 5.1) and word embeddings (Section 5.2).

5.1 Lexical Specificity

Initially, an attempt was made to identify similarities and differences between the European and non-European tweets for each language subset. This was achieved by computing the lexical specificity value of each word. Lexical specificity is a statistical measure which calculates the set of most representative words for a given text based on a reference corpus and the hypergeometric distribution [4, 20]. In contrast to similar scores used to calculate importance of terms, such as TF-IDF, lexical specificity is not especially sensitive to different text lengths and does not require a full partition of the corpus.

Table 5: Top terms, along with their respective lexical specificity score, for the European (E) and non-European (NE) subsets of each language for each year studied.

2019	English E	brexit - 17569	die - 14801	bbc - 14330	electoral - 9389	farage - 5883
	English NE	trump - 19487	mueller - 9453	obama - 7935	media - 7216	president - 7067
	Spanish E	advertencia - 235	esbirros - 176	hecha - 146	terrorista - 130	asesina - 122
	Spanish NE	banco - 204	engañar - 164	presidente - 137	quer - 121	bolsonaro - 109
	French E	faire - 2837	plus - 2355	fait - 2282	macron - 1415	monde - 1245
	French NE	mueller - 1050	trump - 1039	faux- 724	clinton - 501	spécial - 492
2020	English E	tory - 12389	boris - 11499	cummings - 10597	forgotten - 9098	johnson - 8879
	English NE	trump - 13024	president - 8497	obama - 7107	democrats - 4807	election - 3944
	Spanish E	sánchez - 5093	sono - 4463	españa - 3484	gobierno - 3211	vox - 2931
	Spanish NE	trump - 7791	india - 3444	fox - 2457	ccp - 2286	própria - 2258
	French E	plus - 2232	faire - 2226	fait - 1879	meuf - 1639	bien - 1348
	French NE	eua - 1388	sedition - 1224	secession - 1219	ccp - 754	venezuela - 639

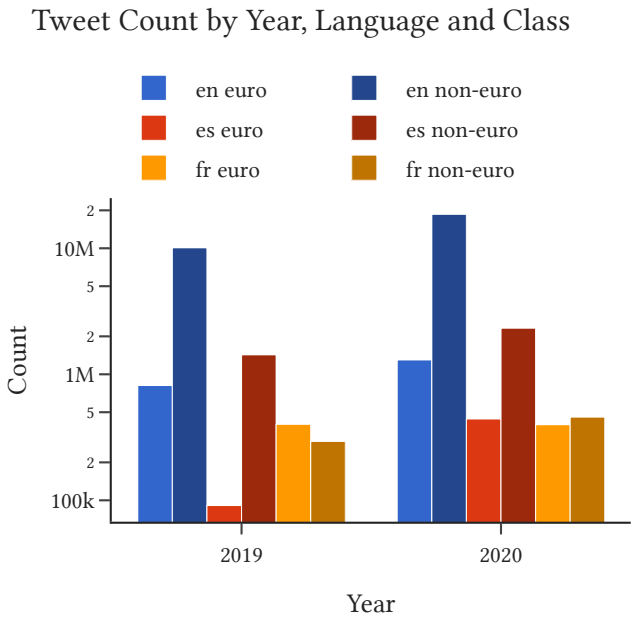

Figure 2: Filtered Tweet Count by Year, Language and Class.

Table 5 displays, for each language, the top five relevant terms according to lexical specificity with respect to the corpus of each year, when considering the European and non-European subsets separately. To gain a better understanding of tweets content, Table 5 does not include words that do not belong to the respective language (e.g. only French words were considered for the French subsets). One interesting observation is that for every language the European and non-European sets appear to have different terms. For example, for the English 2019 subset the European corpus is focused on the topic of Brexit while in the non-European corpus terms were found related to USA politics (e.g., 'trump' and 'obama'). Similarly, when considering the Spanish 2020 subset the European part revolves around Spain with terms like 'sánchez' (Pedro Sánchez being the Spanish prime minister) and 'españa', while the non-European subset seems to be more international with terms like 'ccp', 'india' and 'trump'. These results verify, in a way, that the classification process applied was successful.

Another interesting observation is the almost complete change of topic for the English European corpus from Brexit related terms in 2019 to more generic political ones in 2020. There is also an evolution of the Spanish European corpus from intimidating terms in 2019, such as 'terrorista' (terrorist) and 'esbirros' (thugs), to a more 'nationalistic' turn in 2020 with terms like 'españa' (Spain) and 'gobierno' (government).

5.2 Embeddings

The natural language processing libraries spaCy [13] and gensim [27] are used to preprocess the tweet texts. The extended version of the tweet is used and retweets are included. The text is tokenized and lemmatized with punctuation removed. The 'RT' token present at the start of any retweets as well as any urls are removed. The phrase detection technique introduced by Mikolov et al. [24] is applied to the text with significant bigrams concatenated into a single string delimited by an underscore character. These phrases are considered individual tokens in training.

While pre-trained models have become the foundation to many NLP applications, they are primarily designed to generalize. In this case the latent aspects of interest can be more easily discovered by training a language model using solely the data to be investigated. To achieve this, Word2vec [23] is used with the continuous bag-of-words (CBOW) model architecture to create the embeddings.

English. Table 6 shows the ten most similar words for two queries, 'immigrant' and 'vaccine' for each year and by geographic region in English. For the 'immigrant' query the most striking result is the learned terms for ethnic groups that would be expected to be associated with the geographic region. For example 'greeks' and 'europeans' in the 2020 English European model compared with 'mexicans' and 'blacks' in the English non-European model. There are expected terms mixed in as well such as 'immigration', 'migrant', 'refugee' and 'foreigner'. Other differences include multiple learned terms relating to Judaism ('jews', 'zionists', 'semites') in the 2019 European English set which are not present in the 2020 European

Table 6: The 10 most similar words to the query by year and geographic region for English.

	2019 English			2020 English		
Query	All	European	Non-European	All	European	Non-European
immigrant	migrant	migrant	immigrants	immigrants	refugee	immigrants
	immigrants	semites	immigration	immigration	migrant	immigration
	immigration	immigration	migrant	foreigner	foreigner	foreigner
	refugee	zionists	jews	refugee	migrants	deportation
	jews	refugee	blacks	latinos	greeks	refugee
	mexicans	musli	mexicans	mexicans	europeans	blacks
	quidproquo	suffragette	refugee	asians	pensioner	mexicans
	blacks	jews	invader	invader	settlement	latinos
	invader	vaxer	quidproquo	latino	libyans	asians
	emigrant	semite	labourer	deportation	asians	latino
vaccine	vaccination	vaccination	vaccination	vaccination	vaccination	vaccination
	vaxxers	vape	vaccinations	vaccines	vaccines	vaccines
	vaxer	measles	vaxx	vacine	malaria	vacine
	vaccinations	vaxxers	vaxxers	mmr	cure	mmr
	vape	vaccines	vaxer	medication	tetanus	vac
	vaxx	measle	vape	vac	microchip	rubella
	vaccineswork	tesla	vaxxe	immunization	mmr	medication
	measle	vaxer	measle	microchip	rfid	cure
	vaxxe	mmr	vaccinateyourkids	microchippe	jab	microchip
	vaccinateyourkids	virus	vac	cure	patent	vaxxe

set indicating a shift in the topics. These examples show a clear difference in the use of the word in and outside of Europe in the context of disinformation.

There are also notable differences for the query 'vaccine', particularly to do with conspiracy theories. One of the most popular conspiracies was the assertion that the 2020 Coronavirus Pandemic was a ruse to inject microchips via vaccines. As can be seen in the 2020 English results, 'microchip' and 'rfid' feature in the most similar words to vaccine showing that this method has the ability to identify emerging or trending conspiracies.

Spanish. Table 7 shows the ten most similar words for two queries, 'inmigrante' (immigrant) and 'vacuna' (vaccine) for each year and by geographic region in Spanish. For the query 'inmigrante' (immigrant) the most similar word across all three geographic regions for 2019 is 'perjuicio' (damage/detriment) which suggests that the word is being used in a negative context. For 2020, the top word across all three geographic regions is 'copia' (copy) which initially appears odd. However, on inspecting the data there are multiple retweets about creating a propaganda video for Vox (a far-right Spanish political party) blaming immigrants for selling pirated media.

For the query 'vacuna' (vaccine) there is a clear difference between the two years. The top results 2019 include 'vih' (HIV) and 'vph' (HPV) which mirror common misinformation and disinformation spread by anti-vaxxer groups stating that vaccines result in these illnesses. There are also words that would be expected such as 'inmunización' (immunization), 'vacunación' (vaccination) and gripe (flu) as well as unexpected words such as 'pornografía' (pornography) and 'irak' (Iraq). For the year 2020, there are results more in keeping with what would be expected from a generalized language model mixed in with multiple terms to do with consiracy theories such as 'microchip' and 'bill_gates'. One of the most popular conspiracies was the assertion that the 2020 Coronavirus Pandemic was a ruse to inject microchips via vaccines.

French. Table 8 shows the ten most similar words for two queries, 'immigré' (immigrant) and 'vaccin' (vaccine) for each year and by geographic region in French. For the query 'immigré' (immigrant) the most similar terms for non-European 2019 are 'athmane_tartag' and 'mohamed_médiène' referring to the arrest of two Alegerian intelligence officials. The rest of the results for 2019 are quite mixed with many of the words being related to ideologies or pertain to the ruling of the state for example 'nationalisme' (nationalism), 'république' (republic) and 'nation' (nation). For both years there are terms that suggest a threat such as 'occupation' (occupation), 'invasion' (invasion) and terrorisme (terrorism) which is language common in far-right rhetoric.

For the query 'vaccin' (vaccine) 'big_pharma' appears in reference to a conspiracy theory that states the pharmaceutical industry has malevolent ulterior motives. This is especially relevant as the period is in the beginnings of the 2020 COVID-19 pandemic. 'id2020' is a genuine organisation that provides identification services. Misinformation spread stating that a vaccination program by the organisation and Bill Gates aimed to give people worldwide a digital ID. 'Hydroxycholorquine' and an abbreviation 'hcq' refer to the antimalarial medicine that misinformation categorised as a 'cure' for Coronavirus when in reality it was an experimental treatment.

Table 7: The 10 most similar words to the query by year and geographic region for Spanish.

	2019 Spanish			2020 Spanish		
Query	All	European	Non-European	All	European	Non-European
inmigrante	perjuicio	perjuicio	perjuicio	copia	copia	copia
	embajada	laicidad	inmigracion	mapuch	televisión_sectario	mapuch
	inmigracion	divisa	adve	inmigración	derribo	turista
	laicidad	via	embajada	turista	example	inmigración
	amanecerrcn	prado_miembro	renta	colono	estratagema	example
	backstage	estados	leyva	etnia	sodomía	vivienda
	demócrata	años	cuneta	paguita	fachada	campesino
	republicanos	cultivo	etnia	islam	difamación	sirios
	etnia	inspección	rebelión	gitanos	niña	crer
	manada	descarga	estancamiento	inmigracion	acoso	beneficios
vacuna	vih	vih	vph	vacunación	chip	tratamiento
	vph	live	vih	tratamiento	laboratorio	vacunación
	neumonia	investigación	vacunación	cura	microchip	cura
	anorexia	mod	neumonia	medicamento	medicamento	medicamento
	estigma	auge	gripe	vacunas	bill_gates	microchip
	leaving_neverland	taller	mkt	chip	virus	gripe
	inmunización	ataque	fármaco	microchip	nanochip	virus
	pornografía	irak	virus	virus	sida	inyección
	pastilla	acciones	inmunización	sida	vacunación	vacunas
	vacunación	phishing	musicoterapia	vih	humanidad	chip

Table 8: The 10 most similar words to the query by year and geographic region for French.

	2019 French			2020 French		
Query	All	European	Non-European	All	European	Non-European
immigré	invasion	souche	athmane_tartag	délinquance	colonie	immigration
	arabie_saoudite	pauvreté	mohamed_médiène	colonie	tradition	réfugié
	civil	république	moise	immigration	banlieue	banlieue
	pauvreté	humiliation	abdallah	réfugié	délinquance	ouest
	occupation	résistance	fisc	banlieue	algérie	esclavagisme
	quota	président	nezzar	paysan	souveraineté	occupation
	référendum	travailleur	macky	monarchie	souverain	colonie
	nation	nationalisme	triade	tribu	richesse	délinquance
	résistance	richesse	glyphosate_monsanto	tradition	colonisation	tradition
	traître	invasion	impérialisme	esclavagisme	esclavage	terrorisme
vaccin	vaccination	vaccination	veritable_islam	remède	vaccination	traitement
	glyphosate	lutte	lutte	médicament	puce	médicament
	lutte	généralisation	glorieuse_nation	vaccination	id2020	remède
	méfiance	maladie	noms	puce	chloroquine	puce
	ameriquelatine	élevage	triade	médoc	médoc	virus
	élevage	mobutu	antisémitisme	chloroquine	médicament	hcq
	blanchiment	polio	signataire	bill_gate	bill_gate	big_pharma
	scrat	mutinerie	diatlov	traitement	traitement	bill_gates
	maçonnerie	eglise	populisme	id2020	hcq	vaccination
	généralisation	lyme	élevage	gates	big_pharma	covid

5.2.1 Analogical Reasoning. One of the main benefits of word embeddings, as shown in [23, 24] is the ability to perform analogical reasoning by computing the relational similarity between two word pairs $\langle \mathbf{a}, \mathbf{b} \rangle$ and $\langle \mathbf{c}, \mathbf{d} \rangle$ by finding the most similar word associated with the resulting vector $\mathbf{d}$ (measured usually by cosine distance) to a query consisting on $\mathbf{d} = \mathbf{b} + \mathbf{a} - \mathbf{c}$. For example 'London - Britain

Table 9: Analogical reasoning examples using the English 2019 All Word2vec model (predictions in bold).

English 2019 All			
bbc	britain	**abc**	america
trump	america	**boris_johnson**	britain
politician	government	**md**	hospital

Table 10: Analogical reasoning examples using the English 2020 All Word2vec model (predictions in bold).

English 2020 All			
bbc	britain	**fox**	america
trump	america	**drumpf**	britain
politician	government	**drs**	hospital

+ Spain = Madrid', which in natural language can be phrased as 'London is to Britain as Madrid is to Spain'. In such case, a *capital-of* relationship is learned and revealed via this operation. Table 9 and Table 10 list examples of these arithmetic operations using the English embeddings we use in this paper. The third element on each row is predicted by using the first, second and fourth words.

The first row shows that the 'American' and 'British' qualities of media organizations have been learned with different outlets for 2019 and 2020, 'abc' and 'fox' respectively. In the second row for 2020, the learned analogy is incorrect with 'drumpf' being the original surname of Donald Trump's family. The third row shows a more generic example, with different short forms for 'doctor'.

5.2.2 Disinformation Surrounding the Origins of COVID-19. A particularly successful conspiracy in the English language from early 2020 was that COVID-19 originated in a laboratory. Various flavours of this disinformation circulated ranging from rumours that the virus had been accidentally released to assertions that it was an American or Chinese biological weapon. Table 11 shows the top 5 most similar words to 'laboratory' in the 2019 and 2020 English All models. There is a clear absence of terms relating to this conspiracy in 2019 and the strong presence of it in 2020. Other conspiratorial themes appear in the French and Spanish embeddings models though these are omitted for brevity.

The most similar words for 2019 are mundane terms that are related to the word 'laboratory'. In comparison with 2020, the most similar words relate to this conspiracy including 'wuhan' and 'wuhan_lab' for the Wuhan Institute of Virology, and the United States military lab 'fort_detrick' for the American version. These relate to the United States and Chinese counterparts of these analogous strands of disinformation. These examples show a dramatic change in the use of the term 'laboratory' in the context of disinformation. This finding aligns with other studies, which have used word embeddings to demonstrate semantic shift during the pandemic [11].

Table 11: The 5 most similar words to the query 'laboratory' by year for the model English All.

English All 2019	English All 2020
furniture	lab
warehouses	biolab
rebar	wuhan_lab
extrusion	fort_detrick
shoplife	wuhan

6 CONCLUSION & FUTURE WORK

This paper shows that user-generated content in multiple languages can be used as a data source for deriving insights into disinformation. To achieve this, first a transformer-based classifier is trained on the 0.34% of 87.9 million tweets that contain geolocation data which is then applied to the rest of the data, separating it into European and non-European tweets. This is done for two periods, 2019 and 2020, in English, French and Spanish allowing for multiple types of comparative analysis. It is demonstrated that monolingual classifiers trained and tested on data from the same year outperform multilingual classifiers. Furthermore, it is shown that the geolocation metadata from a relatively small subset of tweets can be used to classify the entire set. An advantage of this method is that the data used to train the classifier is self-contained and usable so long as there is a large enough volume of geolocated tweets to make machine learning methods viable. Secondly, lexical specificity and word embeddings are used to explore the classified tweets and reveal insights into disinformation. For example, it is shown that the conspiracies surrounding the origin of COVID-19 are revealed through comparing the most similar words to a relevant keyword.

Future work could include classifying the data at a lower levels of granularity, for instance at country level by simply using the country code instead of grouping them into broader regions. A popular method of visualising word embeddings is by projecting the vectors into 2 dimensions using a method such as t-SNE [30]. This type of visualisation could form part of an end-to-end system that would allow subject-matter experts with limited technical training to conduct these analyses. Experiments are also being conducted to turn the results of the analytic methods into query and 'dashboard' tools for analysts.

REFERENCES

[1] Jordan Bakerman, Karl Pazdernik, Alyson Wilson, Geoffrey Fairchild, and Rian Bahran. 2018. Twitter Geolocation: A Hybrid Approach. *ACM transactions on knowledge discovery from data* 12, 3 (2018), 1–17.

[2] Francesco Barbieri and Jose Camacho-Collados. 2018. How gender and skin tone modifiers affect emoji semantics in Twitter. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. 101–106.

[3] José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*.

[4] José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence* 240 (2016), 36–64.

[5] Ryan Compton, David Jurgens, and David Allen. 2014. Geotagging one hundred million twitter accounts with total variation minimization. In *2014 IEEE international conference on Big data (big data)*. IEEE, 393–401.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association*

for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[7] Xin Dong and Gerard de Melo. 2019. A Robust Self-Learning Framework for Cross-Lingual Text Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 6306–6310. https://doi.org/10.18653/v1/D19-1658

[8] Ben Eisner, Tim Rocktäschel, Isabelle Augenstein, Matko Bošnjak, and Sebastian Riedel. 2016. emoji2vec: Learning emoji representations from their description. *arXiv preprint arXiv:1609.08359* (2016).

[9] Pedro Henrique Arruda Faustini and Thiago Ferreira Covões. 2020. Fake news detection in multiple platforms and languages. *Expert Systems with Applications* 158 (2020), 113503. https://doi.org/10.1016/j.eswa.2020.113503

[10] Yoav Goldberg. 2017. Neural network methods for natural language processing. *Synthesis lectures on human language technologies* 10, 1 (2017), 1–309.

[11] Yanzhu Guo, Christos Xypolopoulos, and Michalis Vazirgiannis. 2021. How COVID-19 Is Changing Our Language : Detecting Semantic Shift in Twitter Word Embeddings. (2021).

[12] Bo Han, Paul Cook, and Timothy Baldwin. 2014. Text-based twitter user geolocation prediction. *The Journal of artificial intelligence research* 49 (2014), 451–500.

[13] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. *spaCy: Industrial-strength Natural Language Processing in Python.* https://doi.org/10.5281/zenodo.1212303

[14] Binxuan Huang and Kathleen M Carley. 2017. On Predicting Geolocation of Tweets Using Convolutional Neural Networks. In *Social, Cultural, and Behavioral Modeling (Lecture Notes in Computer Science, Vol. 10354)*. Springer International Publishing, Cham, 281–291.

[15] Yin-Fu Huang and Po-Hong Chen. 2020. Fake news detection using an ensemble learning model based on Self-Adaptive Harmony Search algorithms. *Expert Systems with Applications* 159 (2020), 113584. https://doi.org/10.1016/j.eswa.2020.113584

[16] David Jurgens. 2013. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 7.

[17] David Jurgens, Tyler Finethy, James McCorriston, Yi Xu, and Derek Ruths. 2015. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 9.

[18] Soufia Kausar, Bilal Tahir, and Muhammad Amir Mehmood. 2020. ProSOUL: A Framework to Identify Propaganda From Online Urdu Content. *IEEE access* 8 (2020), 186039–186054.

[19] Konstantin Kobs, Albin Zehe, Armin Bernstetter, Julian Chibane, Jan Pfister, Julian Tritscher, and Andreas Hotho. 2020. Emote-Controlled: Obtaining Implicit Viewer Feedback Through Emote-Based Sentiment Analysis on Comments of Popular Twitch.tv Channels. *ACM transactions on social computing* 3, 2 (2020), 1–34.

[20] Pierre Lafon. 1980. Sur la variabilité de la fréquence des formes dans un corpus. *Mots. Les langages du politique* 1, 1 (1980), 127–165.

[21] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised Language Model Pre-training for French. In *Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 2479–2490. https://www.aclweb.org/anthology/2020.lrec-1.302

[22] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[23] Tomas Mikolov, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. http://arxiv.org/abs/1301.3781

[24] Tomas Mikolov, Ilya Sutskever, Kai Chen, G.s Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems* 26 (10 2013).

[25] Pavlos Paraskevopoulos and Themis Palpanas. 2016. Where has this tweet come from? Fast and fine-grained geolocalization of non-geotagged tweets. *Social network analysis and mining* 6, 1 (2016), 1–16.

[26] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[27] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50.

[28] Kai Shi, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Mining Disinformation and Fake News: Concepts, Methods, and Recent Advancements. In *Disinformation, Misinformation, and Fake News in Social Media Emerging Research Challenges and Opportunities* (1st ed. 2020. ed.).

[29] James Thorne and Andreas Vlachos. 2018. Automated Fact Checking: Task Formulations, Methods and Future Directions. In *Proceedings of the 27th International Conference on Computational Linguistics*. 3346–3359.

[30] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[31] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6

[32] Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. *arXiv preprint arXiv:1904.09077* (2019).

[33] Shijie Wu and Mark Dredze. 2020. Are All Languages Created Equal in Multilingual BERT? *arXiv preprint arXiv:2005.09093* (2020).