Classifying Information Sources in Arabic Twitter to Support Online Monitoring of Infectious Diseases

Lama Alsudias
King Saud University / Saudi Arabia
Lancaster University / UK
`lalsudias@ksu.edu.sa`

Paul Rayson
Lancaster University / UK
`p.rayson@lancaster.ac.uk`

Abstract

There is vast untapped potential in relation to the use of social media for monitoring the spread of infectious diseases around the world. Much previous research has focussed on English only, but the Arabic twitter universe has been comparatively much less studied. Motivated by important issues related to levels of trust, quality and reliability of the information online, here we consider the variety of information sources. As a first step, we find that numerous accounts disseminate information via Arabic social media, and we group them into five types of sources: academic, media, government, health professional, and public. We perform two experiments. First, native speakers judge whether they can manually classify tweets into these five groups, and then we repeat the experiment using various Machine Learning (ML) classifiers. We find that inter-annotator agreement is 0.84 for this task, and ML classifiers are able to correctly identify the type of source of a tweet with 77.2% accuracy without knowledge of the user and their bio or profile, but with 99.9% accuracy when provided with this information.

Keywords: Arabic, Infectious diseases, Machine Learning, Natural Language Processing, Twitter.

1 Introduction

People participate in the social web to express their opinions and provide information, which also gives researchers the opportunity to analyse those opinions across larger scale populations than is otherwise possible. One aim of this process is to summarise general opinions regarding international, national, or local events or themes in the huge amount of data available on the Internet. Twitter[1], one of the most popular tools for microblogging in real time, is used by people and organisations alike to share information on different topics, and these can include emergency and/or vital public health information. Due to its popularity as a communication platform, it can be difficult to distinguish reliable information from popular opinions or rumours and this is especially problematic in emergency or health-related scenarios.

In this paper, we consider that it is important to assist reliability and trust judgements by taking into account the source of the information alongside the content of tweets. Hence, via the Twitter API we collected 1,266 tweets containing information about infectious diseases which we categorised into five types of sources: academic, media, government, health professional, and public. First, in order to validate the suitability of the groupings, two Arabic native speakers performed an independent manual labelling of each tweet into one of the types. The resulting inter-coder reliability was 0.84. Second, in order to see whether the grouping can be replicated on a much larger scale, we applied several ML models including Logistic Regression, Random Forest Classifier, Multinomial Naïve Bayes Classifier, and Linear Support Vector classifier. We evaluated the results using 10 fold cross validation for each model. The linguistic features we used to train the systems were selected via the best features based on univariate statistical test. The results show that the ML algorithms correctly classified tweets with up to 77.2% accuracy. We also prove that the bio of the tweet source is an important key that can be used in classification.

[1] http://twitter.com/

The rest of paper is organized as follows. Section 2 covers related background work. Section 3 describes the data collection methods. Section 4 presents the manual coding and Inter-rater reliability. Section 5 focuses on the ML models. Section 6 discusses and visualizes the results. Section 7 contains our conclusions and suggests future work.

2 Related work

There has been a growing interest in Arabic language processing on social media data and Twitter in particular, but only a small part of this research is related to health and medicine topics as we show in this section. In contrast, a growing number of studies have analysed tweets for the public health information they contain in English and other languages such as Chinese and German (Charles-Smith et al., 2015). We summarise this research in what follows.

Arabic related research: Very few papers have described the use of Arabic tweets for studies around health and medicine topics. Khalid and others (2015) evaluated the correctness of health information on twitter based on its medical accuracy. They found that 51.2% of tweets contained false information. The study showed the need for policies on using the social media for health care information. The study of (Alayba et al., 2017) used twitter for sentiment analysis of health services. They collected unbalanced twitter data and annotated it using three annotators by selecting the most frequent tags in each case. They used ML and deep neural networks techniques in their experiment and achieved 91% accuracy using support vector machines. A survey performed by (Alsobayel, 2016) concluded that twitter is the most frequently used by health care professionals in Saudi Arabia for the aim of professional development.

English related research: In contrast to Arabic, there is a much larger body of work utilising twitter in English for research on health and medicine related topics. The previous studies of Paul and Dredze (2011), Aramaki et al. (2011), Breland et al. (2017), and Sinnenberg (2017) have proved that twitter contains valuable information on public health. These studies show the power of Natural Language Processing (NLP) techniques for learning new information from twitter for public health research and to support health informatics hypotheses. The Epidemic Sentiment Monitoring System (ESMOS) is an example of tools that visualise Twitter users' concerns towards specific health conditions developed by Ji, Chun, and Geller (2013) in order to reduce the time of identifying peaks and ongoing monitoring of diseases required by public health officials. They also displayed a knowledge-based approach that utilises a medical ontology, an open source Disease Ontology developed by (Arze et al., 2011), to identify the occurrence of illnesses and to analyse the etymological expressions that provide subjective expressions and polarity of emotions, sentiments, conclusions, individual states of mind, etc. with an opinion classifier (Ji et al., 2016). In (Yepes et al., 2015), pilot results were demonstrated in relation to future directions to investigate when using Twitter information for public health. Other research (Charles-Smith et al., 2015; Paul et al., 2016) has analysed social media articles in supporting public health in order to show the effectiveness of this strategy. These papers recommend a combination of social media with other techniques to measure disease surveillance and spread. The goal of the study in (Sivasankari et al., 2017) was to produce a real time system for the prediction and detection of the spread of an epidemic by identifying disease tweets by graphical location. Good accuracy and quite diverse expressions were uncovered targetting health-related subjects (Doan et al., 2018). Although the results depend only on four months of Twitter data, the paper develops a useful approach to extracting cause-effect relationships from tweets. Other researchers have combined Twitter data with Google Trends data for tracking the spread of infectious diseases (Hong and Sinnott, 2018). A study by Ahmed and others (2018) analysed the twitter data from infectious disease outbreaks. The study developed new insights into how users respond during infectious disease outbreaks and reflects on users' response in association with the sociological concept of the moral panic. They also suggest to examine the tweets in other languages.

Although the above studies use a variety of different approaches in NLP for showing how twitter data can be used for public health applications including monitoring the spread of some diseases, they only consider the information content and do not attempt to study the trust or reliability of the sources of information. While social media users share information about disease outbreaks, symptoms, drug interactions, diet success, and other health behaviors (Paul et al., 2016; Yepes et al., 2015), more than half of the tweets may contain false information (Alnemer et al., 2015). Hence, there is a clear requirement to filter this information in some way, and hopefully to study the ways in which we can reduce the noise of false information and to find tweets with more reliable information of a higher quality. A key first step to do this is to consider the source of the information since this helps the reader to determine how reliable the information is, for example by comparing a tweet on a specific topic by a health professional versus something similar via a member of the general public. Most previous research has focussed on the content of the tweets and not on the source of the information, hence we take a new approach in this paper to combine the two elements together.

3 Data collection

First, we collected 10,000 Arabic tweets via the Twitter API between December 2018 and February 2019. We defined the keywords related to infectious diseases. These keywords were generated by translating an English Disease Ontology, a medical Ontology, developed by Schirmal et al (2011) since we were unable to locate a suitable Arabic equivalent. Using the Disease Ontology allows us to find all terms related to infectious diseases with a set of synonyms (Ji et al., 2016). The Twitter API does not allow us to retrieve enough historical tweets unless we know the user ID in advance. Therefore, we followed two strategies in collecting tweets:

- collect tweets containing words from the keyword list.
- from a suitable user account (discovered using the first step), collect all historical tweets, and filter them depending on the keyword list.

Next, we filtered the tweets manually by removing advertisements, spam, and retweets. After that, we devised Python scripts to clean the tweets by removing URLs, mentions, hashtags, numbers, emojis, repeating characters and non-Arabic words. We also automatically normalised the Arabic tweets and tokenised them. The first author of the paper first classified each of the tweets manually into one of the five groups described in Section 3.1. Second, we asked another Arabic native speaker to independently classify the tweets into the same five groups in order to calculate inter-coder reliability as described in Section 4.2.

3.1 Tweet categorisation

Based on our initial manual reading of the tweets, we decided on five types of users to classify the tweets into, taking into account the various levels of trust that might be associated with each type. For each category listed below there is a small description with examples illustrated in Table 1.

Academic: the tweet is written by academic researchers in higher education. To illustrate, this could be a researcher who carries out studies about infectious diseases.

Media: the tweet is written for newspapers or magazines whether they are general media or health specific ones. In most cases it contains news about infectious diseases.

Government: the tweet is written by a user account that represents the government in some official capacity such as the ministry of health. It may include news, admonition, warning, or general information related to infectious diseases.

Health professionals: the tweet is written by doctors, nurses, or other health service practitioners. In other words, any person who is employed or trained in the health domain and writes the tweet on any information related to infectious diseases.

Public: the tweet is written by members of the general public. It may include information on infectious diseases, feeling sick, or giving advice to someone. Also, it may be written in many dialects since the people come from many Arab countries.

Table 2 shows the number of tweets in each category after filtering and preprocessing. The total number of tweets is 1,266 with only 56 tweets in the media category and 436 tweets in the public one. The reason for this is that there are few media accounts that tweet about infectious diseases whilst many members of the public tweet on the topic. In the other categories (academic, government, and health professionals), there are a relatively well balanced number of tweets which are 239, 258, and 277, respectively.

Category	Tweet in Arabic	Translated tweet to English
Academic	ورقة علمية حديثة تراجع وتتناول فيروس الهربس وتأثيراته الاكلينيكية خاصه على الأطفال ومضى ضعف المناعة .	A recent scientific paper reviews the Herpes virus and its clinical effects on children and patients with immunosuppression.
Media	حاله وفاه بسبب عدوى الكورونا في خميس مشيط .	A case of death due to infection of the corona in Khamis Mushayt.
Government	يتوفر لقاح الانفلونزا في مراكز الرعايه الصحيه الأوليه .	The flu vaccine is available in primary health care centers.
Health Professional	لا تستعمل قطرات للأذن من توصيه مريض سابق ، فعلاج التهابات الاذن الخارجيه البكتيري يختلف عن علاج الالتهاب الفطري .	Do not use ear drops from a previous patient recommendation. Treatment of bacterial external ear infections is different from treatment for fungal infections.
Public	العشبه العجيبه لعلاج أكثر من مرض فى عشبه واحده وهى الزنجبيل .	The wonderful herb to treat more than one disease in one herb is ginger.

Table 1: Examples of tweets in each category

Category	No. of tweets	No. of words
academic	239	3,696
media	56	601
government	258	3,602
health professional	277	6,123
public	436	4,963
total	1,266	18,985

Table 2: Number of tweets and words in each category

4 Manual Coding and Inter-coder Reliability

4.1 Annotation Process

The process starts with labeling the tweets by two Arabic native speakers, including the first author of the paper, who were provided with the guidelines detailed above. The classification depends first on the text in the tweet itself. If the tweet has an ambiguous classification, the annotator may need to look at the bio, a description written by the twitter user, of the user tweet.

4.2 Inter-coder Reliability

We used the Kappa Statistic to test the robustness of the classification scheme (Artstein and Poesio, 2008). The result shows that Cohen's Kappa score is **0.84** which indicates strong agreement between the two manual coders. Figure 1 shows the confusion matrix between the two annotators. The most disparate results between the two coders is between the academic and health professionals categories and this accounts for 10.9% out of the 16% total. This is caused by the similarity between the language of the two groups and the lack of explicit information in the bio to determine which category the tweet fits into.

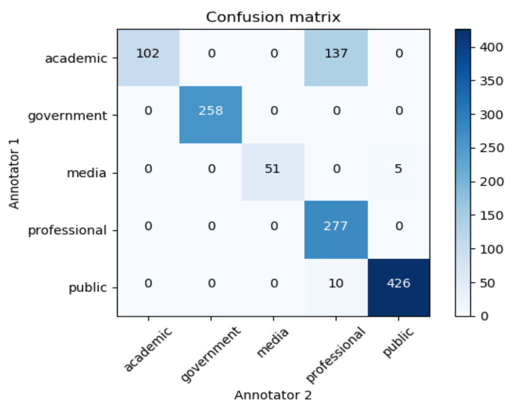

Figure 1: Heat map of confusion matrix between the two annotators.

5 Machine Learning Models

In our study, we used Python Scikit-learn 0.20.2 (Pedregosa et al., 2011) software and applied four ML models: Logistic Regression, Random Forest, Multinomial Naïve Bayes and LSVC: Linear Sup-

Tweet in Arabic	Translated tweet to English	Category
التوصية بأخذ تطعيم فيروس الحصبة للمسافرين .	Recommend vaccination of the measles virus to travelers.	Health professional, government, or academic.
الربيعه : افتتحنا عده مراكز إضافيه لمرضى والاضطرابات وغيرها وتوسعنا في برامج صحه المرأه وخصوصا وأطلقنا مؤشرات لقياس أداء المراكز الصحيه ورضا المراجع .	Alrabiaa: We have opened several additional centers for patients and disorders and others and expanded in women's health programs, especially we have launched indicators to measure the performance of health centers and the satisfaction of references.	Media or government.
بحث لثلاثه أطباء سعوديين يكشف علاقه إنزيم الكبد بالتهاب الزائده الدوديه .	A study by three Saudi doctors reveals the relationship of the liver enzyme to Appendicitis.	Media or academic

Table 3: Examples of tweets with ambiguous categorisation

port Vector Classification. We used 10-fold cross validation to determine accuracy, splitting the entire sample into 90% training and 10% testing for each fold.

5.1 Machine Learning Features

A word frequency approach was used to extract the features from the processed training data after converting them to a matrix of token counts. We designed several features to be used in all four algorithms in order to obtain the best accuracy.

5.1.1 Feature selection

We used various techniques to select the best features automatically (Pedregosa et al., 2011):

- all features: counting unigrams, bigrams and trigrams and ignoring terms that have a document frequency strictly lower than two.
- selecting the best features: Based on a univariate statistical test, keep 60% of the features that have the highest scores.
- selecting from a model: Random Forest Classifier is used to remove unimportant features.
- Using Stanford POST (Part Of Speech Tags) (Manning et al., 2014): Arabic POST is used to annotate the tweet with part-of-speech tags.

5.1.2 Balancing the data

Since the number of tweets is unbalanced across the types, we used two different techniques to re-sample them: under-sampling and over-sampling. We used RUS (Random Undersampling) for under-sampling and ROS (Random Oversampling) for over-sampling. RUS works by removing samples randomly from the majority classes while ROS generates more examples for the minority classes, which has less training data (Burnaev et al., 2015).

5.1.3 Using user bio as a feature

To further resolve the close ambiguity of some types such as academic and health professional, or government and media, we used the bio of the tweet user to provide further features in combination with the tweet text itself. Then we repeated the experiment, including preprocessing, feature extraction and selection, and applied machine learning algorithms, with the new text. Table 3 shows an example of tweets that may be classified into different classes. The first example, which is written by a health professional, can be classified as government because it seems to be providing advice from the government or as an academic as a result from their research.

6 Results and Discussion

Choosing the most frequent class (public) represents 34.4% of the data set, so this represents a simple baseline for our results. Table 4 shows the accuracy, F1-score, Recall, and Precision of the ML models on our training dataset. The Logic Regression classifier and Multinomial Naïve Bayes achieved the highest accuracy (76.0%) with higher recall (0.76) compared to the other algorithms. To evaluate the automatic classification, we compared the algorithms with different sets of features. Figure 2 illustrates the results we achieved running the four ML algorithms with different set of features. The highest accuracy (77.2%) is achieved by the Logic Regression classifier with a selection of the best features based on a univariate statistical test features. It also provided the best score on Random Forest and Linear Support Vector classifications (68.2% and 76.1%, respectively) while selecting all the features reached 76.2% in multinomial NB classifiers. Selecting a model to remove unimportant features did not provide any better results via the four algorithms and POST features had the worst results.

ML	A%	F	R	P
LR	76.0	0.74	0.76	0.74
RF	65.1	0.66	0.68	0.71
MNB	76.0	0.75	0.76	0.75
LSVC	74.1	0.73	0.74	0.75

Table 4: Training results using all features
ML: Machine Learning Model, A%: Accuracy, F: F1-Score, R: Recall, P: Precision
LR: Logistic Regression, RF: Random Forest, MNB: Multinomial Naïve Bayes, LSVC: Linear Support Vector Classification

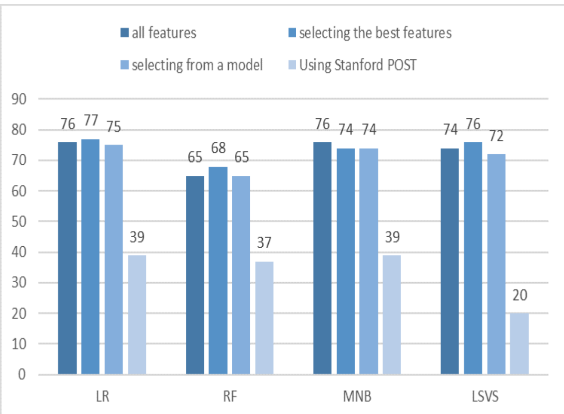

Figure 2: Effect of different features on classification.
LR: Logistic Regression, RF: Random Forest, MNB: Multinomial Naïve Bayes, LSVC: Linear Support Vector Classification

We also compared the normal data, which is the original unbalanced data, with the over and under sampled data to show the effect of re-sampling on the classification accuracy, and the results are

ML	all features	selecting the best features	selecting from a model
LR	76%	79%	77%
RF	62%	72%	72%
MNB	72%	78%	75%
LSVC	73%	78%	76%

Table 5: Effect of over-sampling data on classification.

ML	all features	selecting the best features	selecting from a model
LR	73%	70%	66%
RF	53%	56%	49%
MNB	68%	67%	61%
LSVC	63%	61%	67%

Table 6: Effect of under-sampling data on classification.

shown in Table 5 and Table 6. We can see that there is a small enhancement of accuracy using an over-sampling method especially when selecting the best features and selecting features from a model in all four classifiers. On the other hand, under-sampling the data achieves lower accuracy than normal data due to losing some data in the re-sampling process. Figure 3, Figure 4, and Figure 5 show the comparison between normal data, over-sampling, and under-sampling in all features, selecting the best features and selecting features from a model respectively. We can see that when applying all features, the normal data has the best result in all models expect Logic Regression which has the best result when using over-sampling data with accuracy 77.1% (Figure 3). However, over-sampling data with selecting the best features raises the accuracy between 2% to 4% above normal data in all models (Figure 4). It reaches 79.1% in the Logic Regression classifier and 78.2% in the Multinomial Naïve Bayes and Linear Support Vector Classification models. Moreover, the accuracy is highest when over-sampling data with selecting features from a model in all classifiers for instance, Random Forest classifier which increase from 65.1% to 72.2% (Figure 5).

Table 7 represents the accuracy of each model after combining the bio of the tweet user with the tweet text. In many of the results, the accuracy

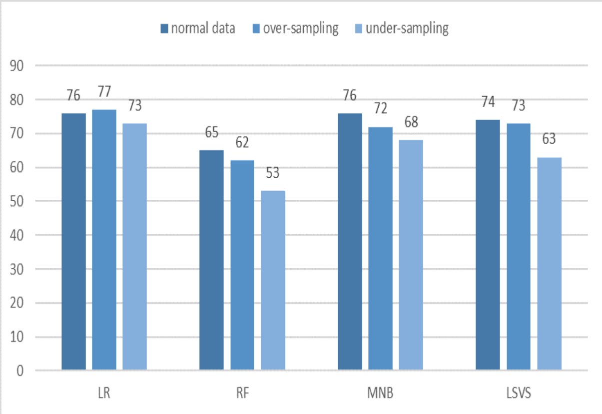

Figure 3: Effect of re-sampling data with all features on classification.

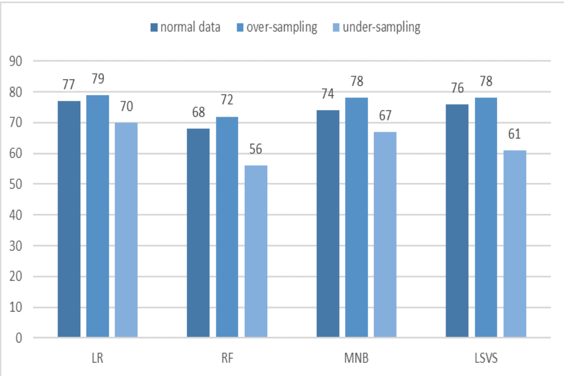

Figure 4: Effect of re-sampling data with selecting the best features on classification.

improves to **99.9%** which shows that the user bio has a very important role to play for classification. For instance, example number 2 in Table 3 can be classified as government after reading the bio of the twitter user which is:
(تتمثل رؤية وزارة الصحة في تحقيق الصحة بمفهومها الشامل على المستويات للفرد والأسرة والمجتمع مع العمل على مساعدة المسنين و ذوي الاحتياجات الخاصة بما يمكنهم.) which means in English: "The vision of the Ministry of Health is to achieve comprehensive health at the individual, family and community levels while working to help the elderly and those with special needs". There are some words in the example bio such as ministry which can help in the classification process. The very high accuracy of the classification can be explained as a result of the limited number of twitter accounts in the study.
We performed an analysis using the Logic Regression classifier with a set of the best features based on univariate statistical test features in the heat map in Figure 6. The matrix shows 12 tweets

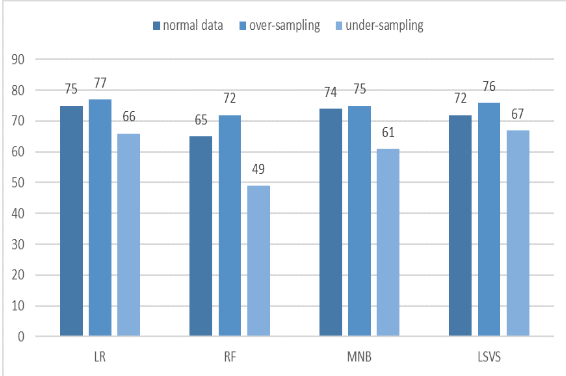

Figure 5: Effect of re-sampling data with selecting from a model on classification.

ML	Tweet text only	Tweet text with bio
LR	77.2%	99.9%
RF	68.2%	99.9%
MNB	76.2%	99.6%
LSVC	76.1%	99.9%

Table 7: Effect of using bio of tweet user as feature on classification.

from the academic category are confused with the government category. In Figure 7 representing the worst accuracy (65.1%) resulting from Random Forest with all features, we assess the degree of confusion between categories. The highest confusion across all categories is between the academic and government types.

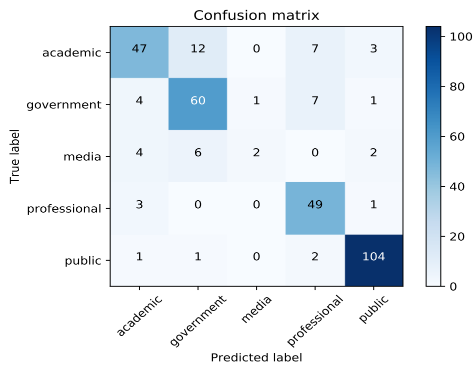

Figure 6: Heat map of confusion matrix of the Logic Regression classifier with selecting the best features based on univariate statistical test features.

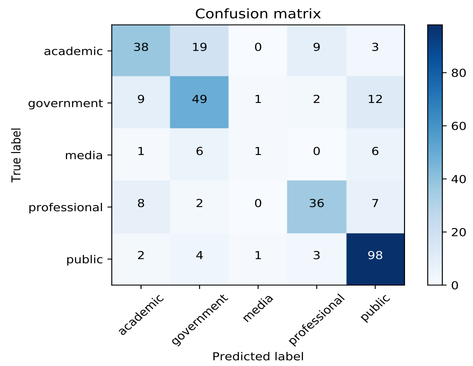

Figure 7: Heat map of confusion matrix of the Random Forest classifier with all features.

7 Conclusion and Future work

In general, social media can be useful as a source of health information. Many government organisations, academic experts, professions, and media outlets in the field of health and medicine release useful health information needed by the public. However, in emergency situations and fast moving scenarios, it is important to understand the veracity of information released in social media, in order to avoid acting on false information. Therefore, we study not only the content of tweets but also the source of the information in order to work towards determining the quality and reliability of public information. We use the bio of the twitter user which contains key information in order to discover reliable accounts that can be trusted.

Here, we introduce a new Arabic social media dataset for analysing tweets related to infectious diseases. The dataset of Arabic tweets has been manually classified into five categories: academic, media, government, health professional, and public, with good inter-rater reliability. We then used ML algorithms to replicate the manual classification. The results showed high accuracy on the classification task, and show that we are able to classify tweets into the five categories with favourable accuracy on the tweet content itself, and highly accurately using the bio information from the user. The dataset, including tweet ID, manually assigned categories, and other resources used in this paper are released freely for academic research purposes[2].

Our future work includes using more NLP techniques and linguistic features such as word embeddings, combining Arabic dialect analysis into the process of classification, and utilising an Arabic medical ontology to be the source of the disease information. Moreover, we will analyse the tweets to support investigation of the spread geographically and over time of infectious diseases in Arab countries.

[2]https://doi.org/10.17635/lancaster/researchdata/303

References

Wasim Ahmed, Peter Bath, Laura Sbaffi, and Gianluca Demartini. 2018. Moral panic through the lens of twitter: An analysis of infectious disease outbreaks. In *Proceedings of the 9th International Conference on Social Media and Society*, pages 217–221. ACM.

Abdulaziz Alayba, Vasile Palade, Matthew England, and Rahat Iqbal. 2017. Arabic language sentiment analysis on health services. In *2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR)*, pages 114–118.

Khalid Alnemer, Waleed Alhuzaim, Ahmed Alnemer, Bader Alharbi, Abdulrahman Bawazir, Omar Barayyan, and Faisal Balaraj. 2015. Are health-related tweets evidence based? review and analysis of health-related tweets on twitter. *Journal of medical Internet research*, 17(10).

Hana Alsobayel. 2016. Use of social media for professional development by health care professionals: A cross-sectional web-based survey. *JMIR Med Educ*, 2(2):e15.

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1568–1576. Association for Computational Linguistics.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Cesar Arze, Gang Feng, Mark Mazaitis, Suvarna Nadendla, Victor Felix, Yu-Wei Wayne Chang, Lynn Marie Schriml, and Warren Alden Kibbe. 2011. Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Research*, 40(D1):D940–D946.

Jessica Breland, Lisa Quintiliani, Kristin Schneider, Christine May, and Sherry Pagoto. 2017. Social media as a tool to increase the impact of public health research. *American Journal of Public Health*, 107(12):1890–1891. PMID: 29116846.

Evgeny Burnaev, Pavel Erofeev, and Artem Papanov. 2015. Influence of resampling on accuracy of imbalanced classification. In *Eighth International Conference on Machine Vision (ICMV 2015)*, volume 9875, page 987521. International Society for Optics and Photonics.

Lauren Charles-Smith, Tera Reynolds, Mark Cameron, Mike Conway, Eric HY Lau, Jennifer Olsen, Julie Pavlin, Mika Shigematsu, Laura Streichert, Katie Suda, et al. 2015. Using social media for actionable disease surveillance and outbreak management: a systematic literature review. *PloS one*, 10(10):e0139701.

Son Doan, Elly W Yang, Sameer Tilak, and Manabu Torii. 2018. Using natural language processing to extract health-related causality from twitter messages. In *2018 IEEE International Conference on Healthcare Informatics Workshop (ICHI-W)*, pages 84–85. IEEE.

Yang Hong and Richard Sinnott. 2018. A social media platform for infectious disease analytics. In *International Conference on Computational Science and Its Applications*, pages 526–540. Springer.

Xiang Ji, Soon Ae Chun, and James Geller. 2013. Monitoring public health concerns using twitter sentiment classifications. In *Healthcare Informatics (ICHI), 2013 IEEE International Conference on*, pages 335–344. IEEE.

Xiang Ji, Soon Ae Chun, and James Geller. 2016. Knowledge-based tweet classification for disease sentiment monitoring. In *Sentiment Analysis and Ontology Engineering*, pages 425–454. Springer.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Michael Paul and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. *Icwsm*, 20:265–272.

Michael Paul, Abeed Sarker, John Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen Smith, and Graciela Gonzalez. 2016. Social media mining for public health monitoring and surveillance. In *Biocomputing 2016: Proceedings of the Pacific Symposium*, pages 468–479. World Scientific.

Fabian Pedregosa, Gaël. Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. 2011. Disease ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40(D1):D940–D946.

Lauren Sinnenberg, Alison Buttenheim, Kevin Padrez, Christina Mancheno, Lyle Ungar, and Raina Merchant. 2017. Twitter as a tool for health research: A systematic review. *American Journal of Public Health*, 107(1):e1–e8. PMID: 27854532.

Si Sivasankari, Mu Kavitha, and Gi Saranya. 2017. Medical analysis and visualisation of diseases using tweet data. *Research Journal of Pharmacy and Technology*, 10(12):4306–4312.

Antonio Yepes, Andrew MacKinlay, and Bo Han. 2015. Investigating public health surveillance using twitter. *Proceedings of BioNLP 15*, pages 164–170.