ISSN 2319-2720
Volume 7, No.2, April – June 2018
International Journal of Computing, Communications and Networking
Available Online at http://www.warse.org/ijccn/static/pdf/file/ijccn36722018.pdf
https://doi.org/10.30534/ijccn/2018/36722018

REAL-TIME CLASSIFICATION OF WORLDWIDE TWEETS AND IT'S FILTERING

Prof. Sunanda V K[1], Sohit D Lotia[2], Sachin Kumar Singh[3], Sapu Ojha[4], Vishal Kumar[5]
[1]Assisstant professor EWIT, India, sunanda@ewit.edu
[2]Student, EWIT, India, sohitdlotia@gmail.com
[3]Student, EWIT, India, sachinkumarsks716@gmail.com
[4]Student, EWIT, India, sappu.ojha@gmail.com
[5]Student, EWIT, India, vk8951327@gmail.com

ABSTRACT

The classification of tweets is done on the basis of using geo-location. With the interest in using social media as the source for research has motivated tackling the challenge of automatically geo-locating tweets. To analyze the tweet's country of origin this can be determined by making use of eight tweet-inherent features and KNN algorithm for classification. Choosing an appropriate combination of both tweet content and metadata can actually lead to substantial improvements . Messages posted on Twitter (tweets) have been reporting everything from daily life stories to the latest local and global news and events. We also make use of Natural Language Processing(NLP) to find out whether the tweet is positive, negative or neutral. We can also know what are the trending hashtags, how many times a tweet is retweeted and we can also perform sentiment analysis by making use of Stanford Core NLP.

Key words : Actionable knowledge, Geo-Location,Metadata,NLP,Tweets**.**

1. INTRODUCTION

Social media are increasingly being used in the scientific community as a key source of data to help understand diverse natural and social phenomena, and this has prompted the development of a wide range of computational data mining tools that can extract knowledge from social media for both post-hoc and real time analysis. Thanks to the availability of a public API that enables the cost-free collection of a significant amount of data, Twitter has become a leading data source for such studies.

Having Twitter as a new kind of data source, researchers have looked into the development of tools for real-time trend analytics[13], or early detection of newsworthy events [4], as well as into analytical approaches for understanding the sentiment expressed by users towards a target [6], public opinion on a specific topic [5]. This has motivated a growing body of research in recent. Most of the previous research in inferring tweets Geolocation has classified tweets by location within a limited geographical area or country; these cannot be applied directly to an unfiltered stream where tweets from any location or country will be observed. The few cases that have dealt with a global collection of tweets have used an extensive set of features that cannot realistically be extracted in a real-time, streaming context (e.g., user tweeting history or social networks) [14], and have been limited to a selected set of global cities as well as to English tweets.

The classifier built on this pre-filtered dataset may not be applicable to a Twitter stream where every tweet needs to be geolocated. This means just using ground truth labels to pre-filter tweets originating from other regions and/or written in languages other than English.. An ability to classify tweets by location in real-time is crucial for applications exploiting social media updates as social sensors that enable tracking topics and learning about location-specific trending topics, emerging events and breaking news. Specific applications of a real-time, country-level tweet geolocation system include country-specific trending topic detection.

To analyze the tweet's country of origin this can be determined by making use of eight tweet-inherent features and KNN algorithm for classification. Choosing an appropriate combination of both tweet content and metadata can actually lead to substantial improvements. Messages posted on Twitter (tweets) have been reporting everything from daily life stories to the latest local and global news and events.

2. RELATED WORKS

A growing body of research deals with the automated inference of demographic details of Twitter users [25]. Digging more deeply into the demographics of Twitter

users,other researchers have attempted socioeconomic demographics such as occupational class [05], income[22] and socioeconomic status [15]. This work is different from that which we report here in that the country where the tweets were posted from, was already known.

3. DATASETS

For training our classifier, we rely on the collection of a Twitter dataset with tweets categorized by location. This involves using the Twitter API endpoint that returns a stream of geolocated tweets posted from within one or more specified geographic bounding boxes.. In our study, we set this bounding box to be the whole world. in order to retrieve tweets worldwide. This way, we collected streams of global geolocated tweets for two different week long periods: 4-11 October, 2014 (TC2014) and 22-28 October, 2015 (TC2015). This led to the collection of 31.7 million tweets in 2014 and 28.8 million tweets in 2015, which we adapt for our purposes as explained below. Our raw datasets reflect the well-known fact that some Twitter users are far more prolific than others, which would introduce a bias in the evaluation if not dealt with. More than 5 million tweets in these two datasets are categorised into 217 different countries. It is worth mentioning that, as one would expect, the resulting datasets are clearly imbalanced, where only a few countries account for most of the tweets. determine the location for users with an empty location field, we default GeoNames' prediction for those tweets to be the majority country, i.e., the United States.

3.1 PROBLEM STATEMENT

Twitter data lacks reliable demographic details that would enable a representative sample of users to be collected and/or a focus on a specifics user subgroup. Most of the previous research in inferring tweet geo-location has classified tweets by location within a limited geographical area or country; these cannot be applied directly to an unfiltered stream where tweets from any location or country will be observed.

The few cases that have dealt with a global collection of tweets have used an extensive set of features that cannot realistically be extracted in a real-time, streaming context (e.g., user tweeting history or social networks) been limited to a selected set of global cities as well as to English tweets.

4. CLASSIFICATION TECHNIQUES

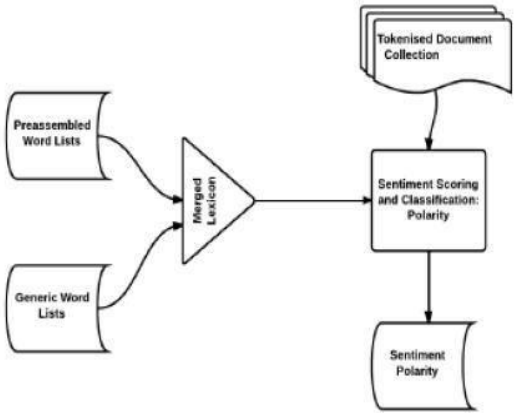
We carried out the experimentation with different classifiers: KNN algorithm, Support Vector Machines (SVM), Gaussian Naive Bayes, Multinomial Naive Bayes, Decision Trees, Random Forests and a Maximum Entropy classifier. They were tested in two different settings, one without balancing the weights of the different classes and the other by weighing the classes as the inverse of their frequency in the training set; the latter was tested as the means for dealing with the highly imbalanced data. The selection of these classifiers is in line with those used in the literature, especially with those tested by Han et al. [10]. This experimentation led to the selection of the KNN algorithm classifier as the most accurate. In the interest of space and focus, we only present results for this classifier. Additionally, we compare our results with two baseline approaches. On the other hand, we used the Vowpal.

5. SENTIMENT ANALYSIS

Lexicon-Based Sentiment Analysis

It uses sentiment dictionary with opinion words and matches them with data. Next it assigns sentiment scores to the opinion words describing it as Positive, Negative and Neutral tweet.

Lexicon-based approaches mainly rely on a sentiment lexicon, i.e., a collection of known and precompiled sentiment terms, phrases and even idioms, developed for traditional genres of communication, such as the Opinion Finder lexicon.



There are two sub classifications for this approach:

1. **Dictionary-Based:**

It is based on the usage of terms (seeds) that are usually collected and annotated manually. This set grows by searching the synonyms and antonyms of a dictionary. An example of that dictionary is WordNet, which is used to develop a thesaurus called SentiWordNet.

Drawback : Can**'**t deal with domain and context specific orientations.

2. **Corpus-Based:**

The corpus-based approach have objective of providing dictionaries related to a specific domain. These dictionaries are generated from a set of seed opinion terms that grows through the search of related words by means of the use of either statistical or semantic techniques.

- Methods based on statistics: Latent Semantic Analysis (LSA).
- Methods based on semantic such as the use of synonyms and antonyms or relationships from thesaurus like Word Net may also represent an interesting solution.

According to the performance measures like precision and recall, we provide a comparative study of existing techniques for opinion mining, including machine learning, lexicon-based approaches, cross domain and cross-lingual approaches, etc., as shown in Table 1.

Table 1. Performance Comparison Of Sentiment Analysis Method

	Method	Data Set	Acc.	Author
Machine Learning	SVM	Movie reviews	86.40%	Pang, Lee[23]
	CoTraining SVM	Twitter	82.52%	Liu[14]
	Deep learning	Stanford Sentiment Treebank	80.70%	Richard[18]
Lexical based	Corpus	Product reviews	74.00%	Turkey
	Dictionary	Amazon's Mechanical Turk	---	Taboada[20]
Cross-lingual	Ensemble	Amazon	81.00%	Wan,X[16]
	Co-Train	Amazon, ITI68	81.30%	Wan,X.[16]
	EWGA	IMDb movie review	>90%	Abbasi,A.
	CLMM	MPQA,NTCIR,ISI	83.02%	Mengi
Cross-domain	Active Learning	Book, DVD, Electronics, Kitchen	80% (avg)	Li, S
	Thesaurus			Bollegala[22]
	SFA			Pan S J[15]

6. EXPERIMENT SETTING

We created eight different classifiers, each of which used one of the following eight features available from a tweet as retrieved from a stream of the Twitter API:

1) User location (uloc): This is the location the user specifies in their profile. While this feature might seem a priori useful, it is somewhat limited as this is a free text field that users can leave empty, input a location name that is ambiguous or has typos, or a string that does not match with any specific

locations (e.g., **"**at home**"**). Looking at users**'** self-reported locations, Hecht et al. [16] found that 66% report information that can be translated, accurately or inaccurately, to a

geographic location, with the other 34% being either empty

or not geolocalisable.

2) User language (ulang): This is the user**'**s self-declared user interface language.

The interface language might be indicative of the user**'**s country of origin; however, they might also have set up the interface in a different language, such as English, because it was the default language when they signed up or because the language of their choice is not available.

3) Timezone (tz): This indicates the time zone that the user has specified in their settings, e.g., **"**Pacific Time (US & Canada)**"**.

When the user has specified an accurate time zone in their settings, it can be indicative of their country of origin; however, some users may have the default time zone in their settings, or they may use an equivalent time zone belonging to a different location (e.g., **"**Europe/London**"** for a user in Portugal). Also, Twitter**'**s list of time zones does not include all countries.

4) Tweet language (tlang): The language in which a tweet is believed to be written is automatically detected by Twitter. It has been found to be accurate for major languages, but it leaves much to be desired for less widely used languages.

5) Twitter**'**s language identifier has also been found to struggle with multilingual tweets, where parts of a tweet are written in different languages [1-25].

6) User description (description): This is a free text where a user can describe themselves, their interests, etc.

7) Tweet content (content): The text that forms the actual content of the tweet. The use of content has a number of caveats. One is that content might change over time, and therefore new tweets might discuss new topics that the classifiers have not seen before. Another caveat is that the content of the tweet might not be location-specific; in a previous study, found that the content of only 289 out of 10,000 tweets was location-specific.

7. EVALUATION

We report three different performance values for each of the experiments: micro-accuracy, macro-accuracy and mean squared error (MSE). The accuracy values are computed as the result of dividing all the correctly classified instances by

all the instances in the test set. The micro-accuracy is computed for the test set as a whole. For macro-accuracy, we compute the accuracy for each specific country in the test set,

which are then averaged to compute the overall macroaccuracy. While the micro-accuracy measures the actual accuracy in the whole dataset, the macro-accuracy penalizes the classifier that performs well only for the majority classes and rewards, instead, classifiers that perform well across multiple categories. This is especially crucial in a case like ours where the categories are highly imbalanced.

We report three different performance values for each of the experiments: micro-accuracy, macro-accuracy and mean squared error (MSE). The accuracy values are computed as the result of dividing all the correctly classified instances by

all the instances in the test set. The micro-accuracy is computed for the test set as a whole. For macro-accuracy, we compute the accuracy for each specific country in the test set,

which are then averaged to compute the overall macroaccuracy. While the micro-accuracy measures the actual accuracy in the whole dataset, the macro-accuracy penalizes the classifier that performs well only for the majority classes and rewards, instead, classifiers that perform well across multiple categories. This is especially crucial in a case like ours where the categories are highly imbalanced. The MSE is the average of the squared distance in kilometres between the predicted country and the actual, ground truth

country, as shown in Equation 1.

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2 \quad (1)$$

and latitude is computed In this computation, the distances between pairs of countries were calculated based on their centroids. We used the Countries of the World (COW) dataset produced by OpenGeonames.org to obtain the centroids of all countries. Having the latitude and longitude values of the centroids of all these countries, we then used the Haversine formula [1-25], which accounts for the spheric shape when computing the distance between two points and is often used as an acceptable approximation to compute distances on the Earth. The Haversine distance between two points of a sphere.
each defined by its longitude as shown in Equation 2.

$$d = 2r\arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2-\varphi_1}{2}\right)+\cos(\varphi_1)\cos(\varphi_2)\sin^2\left(\frac{\lambda_2-\lambda_1}{2}\right)}\right) \quad (2)$$

8. CLASSIFICATION RESULTS

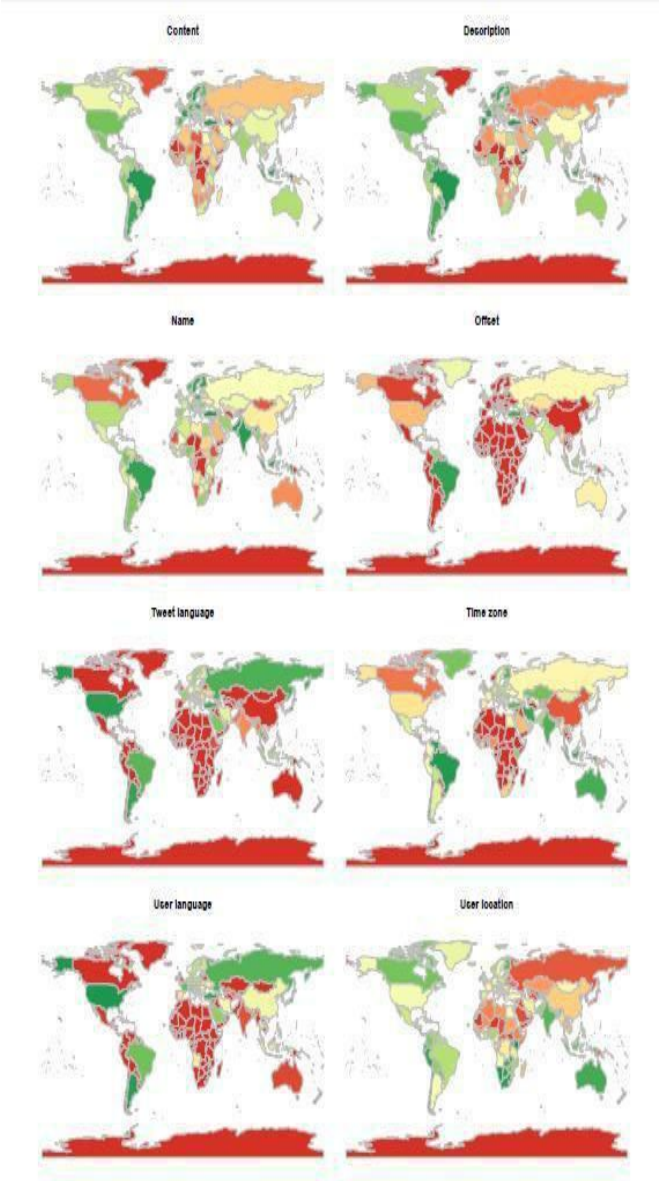
In this section, we present results for different location classification experiments. First, we look at the performance of classifiers that use a single feature. Then, we present the results for classifiers combining multiple features.

8.1 SINGLE FEATURE

There is no big difference between the two approaches based on Geo Names when we look at micro-accuracy. However ,this accuracy is slightly better distributed across countries when we use the approach based on relevance, as can be seen
from the macro-accuracy values. If we look at the micro-accuracy scores, the results suggest that three approaches stand out over the rest. These are tweet content, tweet language and user language, which are the only three approaches to get a micro-accuracy score above 0.5. However, these three approaches leave much to be desired

when we evaluate them based on macro-accuracy scores, and therefore they fail to balance the classification well. Instead, the users' self-reported location (user location) achieves the

highest macro-accuracy scores, while micro accuracy scores are only slightly lower. This is due to the fact that the classifier that only uses the user's profile location will be able to guess location is the only feature to beat the baseline in terms of macro-accuracy. However, the small improvement over the baseline suggests that alternative approaches are needed for a better balanced classification performance .Figure 3 shows a heat map with accuracy values of each of the features broken down by country. We observe the best distributed accuracy across countries is with the use of user location as a feature. However, other features are doing significantly better classifying tweets that belong to some of the major countries such as the USA (better classified by tweet language or user language),Russia.



8.2 FEATURE COMBINATION

Having seen that different features give rise to gains in different ways, testing the performance of combinations of multiple features seemed like a wise option. We performed these combinations of features by appending the vectors for each of the features into a single vector. We tested all 255 possible combinations using the eight features under study. We only report the best performing combinations here in the interest of space and clarity. Table 5 shows the best combination in each case for the TC2014 and TC2015 datasets, as well as for the classifiers that consider all the countries in the datasets and only the top 25 countries.

The table also shows the performance of the best single feature as well as the baseline classifier by [1-25] to facilitate comparison, as well as the improvement in performance when using a combination of features over that of a single feature. We observe that the selection of an appropriate combination of features can actually lead to a substantial increase in terms of all micro-accuracy, macro accuracy and MSE.

All countries							
TC2014				TC2015			
Feature	Micro.	Macro.	MSE	Feature	Micro.	Macro.	MSE
Dredze et al. [14]	0.666	0.122	862.792	Dredze et al. [14]	0.636	0.116	956.997
Best single feature	0.568	0.374	1156.279	Best single feature	0.588	0.370	1148.264
content-description-name-tlang-tz-ulang-uloc	**0.889**	**0.452**	**244.106**	content-description-name-tlang-tz-ulang-uloc	**0.893**	**0.456**	**243.124**
Improvement	+56.5%	+20.9%	-78.9%	Improvement	+51.9%	+23.2%	-78.9%
Top 25							
TC2014				TC2015			
Feature	Micro.	Macro.	MSE	Feature	Micro.	Macro.	MSE
Dredze et al. [14]	0.651	0.513	840.025	Dredze et al. [14]	0.619	0.480	913.611
Best single feature	0.632	0.547	926.810	Best single feature	0.667	0.587	838.722
content-description-name-tlang-tz-ulang-uloc	**0.849**	**0.858**	**360.856**	content-name-tlang-tz-ulang-uloc	**0.837**	**0.853**	**385.807**
Improvement	+34.3%	+56.9%	-61.1%	Improvement	+25.5%	+45.3%	-54.0%

TABLE 5

Results for combinations of features, best performing single feature and the baseline classifier by Dredze et al. [14].

markable when we look at the MSE scores, where the improvement is always above 50%. Improvements in terms of micro-accuracy and macro-accuracy scores are also always above 20%, but are especially high for micro-accuracy (50%+) when we classify for all the countries, and for macroaccuracy (40%+) when we classify for the top 25 countries. These results suggest that the use of a single feature, as it is the case with most previous work using e.g. only tweet content, can be substantially improved by using more features. In fact, our results suggest that the combination of many features is usually best; we need to combine seven of the eight features (all but offset) in three of the cases, and six features in the other case (all but description and offset). As a result, we get performance values above 85% in terms of macro-accuracy for the top 25 countries. These performance scores are also remarkably higher than those of the classifier by [88], both in terms of micro- and macro-accuracy. Interestingly, the combination of features has led to a significant improvement in performance, with a better balance across countries. To complement this analysis, we believe it is important to understand the differences among countries. Will different sets of features be useful for an accurate classification for each country? Are we perhaps doing very well for some countries with certain combinations, but that combination, is in turn, bad for other countries? To explore this further, we now take a closer look at the performance broken down by country.

CONCLUSION

To the best of our knowledge, this is the first study performing a comprehensive analysis of the usefulness of tweet inherent features to automatically infer the country of origin of tweets in a real-time scenario from a global stream of tweets written in any language. Most previous work focused on classifying tweets coming from a single country and hence assumed that tweets from that country were already identified. Where previous work had considered tweets from all over the world, the set of features employed for the classification included features, such as a user's social network, that are not readily available within a tweet and so is not feasible in a scenario where tweets need to be classified in real-time as they are collected from the streaming API. Moreover, previous attempts to geolocate global tweets tended to restrict their collection to tweets from a list of cities, as well as to tweets in English; this means that they did not consider the entire stream, but only a set of cities,which assumes prior preprocessing. Finally, our study uses two datasets collected a year apart from each other, to test the ability to classify new tweets with a classifier trained on older tweets. Our experiments and analysis reveal insights that can be used effectively to build an application that classifies tweets by country in real time, either when the goal is to organise content by country or when one wants to identify all the content posted from a specific country.

In the future we plan to test alternative cost-sensitive learning approaches to the one used here, focusing especially on collection of more data for under-represented countries, so that the classifier can be further improved for all the countries. Furthermore, we plan to explore more sophisticated approaches for content analysis, e.g. detection of topics in content (e.g. do some countries talk more about football than others?), as well as semantic treatment of the content.We also aim to develop finer-grained classifiers that take the output of the country-level classifier as input.

ACKNOWLEDGEMENT

We wish to offer our sincere gratitude to our principal Dr.Prof. K Channakeshavalu, Principal, EWIT, Bangalore, for his moral support towards completing my project work. I would like to thank Dr. Arun Biradar, Head of Department, Computer Science & Engineering, EWIT, Bangalore, for his valuable suggestions and expert advice. I deeply express my sincere gratitude to my guide Prof. Sunanda V K, Assisstant professor Department of CSE, EWIT, Bangalore, for her able guidance throughout the project work and guiding me to organize the report in a systematic manner. I thank my Parents, and all the Faculty members of Department of Computer Science & Engineering for their constant support and encouragement.

REFERENCES

[1] O. Ajao, J. Hong, and W. Liu. A survey of location inference techniques on twitter. Journal of Information Science, 1:1–10, 2015. https://doi.org/10.1177/0165551515602847

[2] E. Amig´ o, J. C. De Albornoz, I. Chugur, A. Corujo, J. Gonzalo,T. **Mart´ın, E. Meij, M. De Rijke, and D. Spina.** Overview ofreplab 2013: Evaluating online reputation monitoring systems. InProceedings of CLEF, pages 333–352. Springer, 2013.

[3] F. Atefeh and W. Khreich. A survey of techniques for event detection in twitter. Computational Intelligence, 31(1):132–164, 2015. https://doi.org/10.1111/coin.12017

[4] H. Bo, P. Cook, and T. Baldwin. Geolocation prediction in social media data by finding location indicative words. In Proceedings of COLING, pages 1045–1062, 2012.

[5] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella. Discriminating gender on twitter. In Proceedings of EMNLP, pages 1301–1309, 2011.