Bi-lingual Intent Classification of Twitter Posts: A Roadmap

Akinlolu Solomon Adekotujo[1,2,3(✉)], JooYoung Lee[2], Ayokunle Oluwatoyin Enikuomehin[1], Manuel Mazzara[2], and Segun Benjamin Aribisala[1]

[1] Department of Computer Science, Lagos State University, Lagos, Nigeria
adekotujoakinlolu@gmail.com, toyinenikuomehin@gmail.com, benjamin.aribisala@gmail.com
[2] Innopolis University, Innopolis, Russia
{j.lee,m.mazzara}@innopolis.ru
[3] Computer, Information and Management Studies Department, The Administrative Staff College of Nigeria, Badagry, Nigeria

Abstract. A core advantage of social media platforms is the freedom that comes with the way users express their opinions and share information as they deem fit, in line with the subject of discussion. Advances in text analytics have allowed researchers to adequately classify information expressed in natural language text, which emanates in millions per minute, under well-defined categories like "hate" or "radicalized" content which provide further insight into intent of the sender. This analysis is important for social media intelligence and information security. Commercial intent classifications have witnessed several research attentions. However, social intent classification of topics in line with hate, radicalized posts, have witnessed little research effort. The focus of this study is to develop a roadmap of a model for automatic bilingual intent classification of hate speech. This empirical model will involve the use of bi-gram words for intent classification. The feature extraction will include expected cross entropy, while topic modeling will use supervised context-based n-gram approach. Classification will be done using ensemble-based approach which will include the use of Naïve Bayes and Support Vector Machine. This study will also discuss the differences between the concept of fake news, stance and intent identification. We anticipate that the proposed roadmap, if implemented, will be useful in the classification of intent as it relates to hate speech in bilingual twitter post. The proposed model has the potential to improve intent classification and that could be useful in hate speech detection, which can avert social or security problems.

Keywords: Intent classification · Hate speech · Machine learning classifier

1 Introduction

Personal expressions or comments on social media are largely about users' emotions, sentiments or goals (intents) which are particularly valuable, for instance, for monitoring activities to ensure security of lives and properties [1]. Understanding user's

intent for hate from speech or text in general, is a natural language problem which is usually difficult to solve, as we are now confronted with much more short texts and news every day. There are three commonly confused terminologies related to personal expressions, opinions, sentiments and emotions, these are fake news, stance and intent. Fake news are intentional false information [2] and stance are the stand of a person on a topic [3] while intent are every day or futuristic behavior or goals. Fake news is commonly motivated by financial and ideological reasons [4]. On the other hand, stance are motivated by sentiments powered by prior knowledge and analytical skills [3], while intent are motivated by individual desires. Intent can help identify actionable information [5, 6].

1.1 Fake News Detection

Fake news detection has recently attracted growing interest in research due to increasing misinformation essentially on social media feeds. Until recently, detection of fake news relied on satire based news system and fact finding news portals like politifact (https://www.politifact.com/). Fake news can be detected by first formalizing the news as an input to a fake news detection equation (Eq. 1) whose output is either 1or 0 (True or False respectively). Given the social news engagements $\mathcal{E}$ among n users for news article a, the task of fake news detection is to predict whether the news article a is a fake news piece or not,

i.e., $\mathcal{F}: \mathcal{E} \rightarrow \{0,1\}$

such that,

$$\mathcal{F}(a) = \begin{cases} 1, \text{ if a is a piece of fake news,} \\ 0, \text{ otherwise.} \end{cases} \quad (1)$$

where $\mathcal{F}$ is the prediction function.

Note that we define fake news detection as a binary classification problem because fake news is essentially a distortion on information released by the publisher [4].

1.2 Stance Detection

Stance detection are also becoming common in research environment. The main aim of stance detection is to identify if the author of a piece of text is supportive of a given target or against it. For clarity, stance should be seen as a subtask of opinion mining. Stance can detection can be represented using Eq. 2, which shows that the output of stance detection model is either Agree (0), Disagree (1), Discuss (2) or Unrelated (3) [7, 8]. Given a set of social media posts D related to the K target entities $T_1,\ldots, T_k$ then, the goal is to determine the value of mapping S. The task of stance detection is to identify the cognitive position of an author of individual posts in his reaction, towards a given statement or claim.

i.e., S: $T_1 \times ... \times T_k \times D \rightarrow \{\text{Agree, Disagree, Discuss, Unrelated}\}$ k for each post $d \in D$

such that,

$$S(d) = \begin{cases} 0, \text{ if agreed to a claim in D,} \\ 1, \text{ if disagree to a claim in D,} \\ 2, \text{ if a claim in D is discussed} \\ 3, \text{ if a claim in D is unrelated.} \end{cases} \quad (2)$$

where S is the claim function.

Stance classification systems normally require identification of a claim as belonging to any of 4 [7, 9–11] categories namely supporting, denying, querying, and commenting. Supporting means that the claim is supported, while denying means that the claim is not supported, but disagreed. Querying implies that the claim is being discussed or questions related to the claim are being raised, while commenting implies that reactions to the claim are unrelated to the claim [3]. Table 1 gives an example of a claim and the reactions related to these four categories [7, 10].

Table 1. A sample claim and reactions showing the four categories of stance

Claim: Robert Plant Ripped up $800M Led Zeppelin Reunion Contract	
Snippet	Stance
Led Zeppelin's Robert Plant turned down £500m to reform supergroup…	Agree
Robert Plant's publicist has described as "rubbish" a Daily Mirror report that he rejected a £500m Led Zeppelin reunion…	Disagree
Robert Plant reportedly tore up an $800 million Led Zeppelin reunion deal…	Discuss
Richard Branson's Virgin Galactic is set to launch SpaceShipTwo today….	Unrelated

1.3 Intent Detection

Intent(ion) detection and analysis are very important aspect of social media modelling. This is largely because intents are hidden and execution of same could unrest or social misbehaviour. Intent detection is commonly used alongside Slot Filling. Using slot filling, the conservational flow within an intent can be determined could serve as a means of determining the validity of the detected intent.

Intent can be detected using Eq. 3. The output of an intent detection algorithm is either True (1) or False (0). Given the social news engagements $\mathcal{E}$ among n users for news article or twitter post a, the task of intent detection is to predict whether the news article or twitter post a has intent of hate speech or not,

i.e., $I: \mathcal{E} \rightarrow \{0,1\}$

such that,

$$\mathrm{I}(a) = \begin{cases} 1, & \text{if a contains hate intent or speech,} \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

where I is the prediction function.

A variety of factors like personal interest, religion and social influence, affect an individual's expression of intentionality or post [12–15]. Intentions that have multiple potential often complicates natural language clarification in short-text documents. In a bid to make the hate intent or goal mining problem computationally manageable, we need to first mine specific intent classes with corresponding hate speech, and therefore define a multiclass intent classification problem.

The task intent analysis is defined [6] as to approximate the unknown function in Eq. (4)

Given that: $S_i \in d_i$, $d_i \subset D$,

$$f: S \times C^I \rightarrow \{True,\ False\}, \tag{4}$$

where $C^I = \left\{c_1^I, c_2^I \ldots c_n^I\right\}$ is a set of predefined intent categories from document, where D is a domain that consist of text documents and each reviewed document d_i contains a sequence of sentences $S = \{s_1, s_2 \ldots s_{|S|}\}$.

Intent classification, is known to focus on futuristic action, it is a form of text classification, whereas sentiment and opinion analysis that compute subjectivity text classification, focuses on the current state of affairs [1, 6]. For instance, in a message "I wanna visit the magnificent Walt Disney animal hotel", topic classification focuses on the noun, the hotel 'Walt Disney Animal Hotel'; sentiment and emotion classification focused on the positive feeling of the author's message expressed with the adjective 'magnificent'. In contrast, intent classification concerns the author's intended future action or goal, i.e. going to visit the hotel. Given that the focus of this study is on intent classification, the next section and subsequent ones will focus on intent classification.

2 Related Work to Intent Classification

This section presents the review of existing works on intent classification. The review was divided into commercial intent classification and social intent classification for easy understanding [12].

2.1 Commercial Intent Classification

Commercial intents are intent that are related to buying and selling or marketing, they can be identified by verbs in a post. Hollerit et al. [16] proposed an automatic method for classifying commercial intent in tweets based on Bayes Complement Naïve Bayes classifier in the textual domain, and a linear logistic regression classifier. Benczúr et al. [17] proposed features for web spam filtering based on the presence of keywords with high advertisement value, same spam filtering power was demonstrated on online commercial intention value.

Lewandowski et al. [18] concluded that neither crowdsourcing approach nor questionnaire approach lead to satisfying result during the survey on a commercial search engine's portal. Guo et al. [19] explore client-side instrumentation for inferring personalized commercial intent of user's, by investigating whether mouse movement over search result can provide clues into the user's intent using clich-through on ads, the result came out that mouse movement analysis can provide such clues.

Lewandowski [20] used retrieval effectiveness test design on three prominent web search engines by classification of the result in relation to their commercial intent, the result showed that Google more significant commercial intent.

Purohit et al. [5] present a hybrid feature approach of combining top-down processing using a bag-of-tokens model to address the problem of multiclass classification of intent on twitter for crisis events dataset and address the problem of ambiguity and sparsity in order to classify the intent of narrative.

2.2 Social Intent Classification

Social intent are intents that do not have verbs representing buying and selling, but that are related to social activities or events. Hate intent is a good example of social intent and the focus of this review is on hate intent.

Ben-David et al. [21] using longitudinal multimodal and network analysis, claimed that hate speech and discriminatory engagements are not only explicated with user's motivations and actions, but also included the formation of a network of ties of platform policy, technological availability, and the manners of communication of its users. The study affirmed that the platform encourages discrimination, which the users eventually exhibits. Wang et al. [22] affirmed that word order and phrases are important in giving understanding to text in many texts mining job, the study proposed a topical n-gram topic model that is able to identify topics and topical phrases using probabilistic model.

The problem of identifying hatred videos is proposed, [23] with the implementation of classification algorithm named shark search, the study focused on the creation of a web portal and produced a framework that will resolve the problem of finding out hatred videos. Agarwal et al. [24], claimed that just a keyword spotting based techniques cannot accurately identify the intent of a post, the study developed a monolingual cascaded ensemble learning classifier for identifying the posts having racist or radicalized intent with the use of open sources API's for feature extractions. It was also reaffirmed [25–28], while presenting an unsupervised method for polarity classification in twitter that little work has been noticed in the area of multilingual domain, the method [25] is based on the expansion of the concepts expressed in the tweets through the application of Page Ranking to Wordnet. Additionally, Gomes et al. and Agarwal et al. [24, 29] emphasized the high performance of ensemble learning for data stream classification of intent above one single classifier. Gomes et al. [29] proposed taxonomy for data stream ensemble learning and listed popular open source tools, while Agarwal et al. [24], used cascaded ensemble learning classifier for only mono lingual post.

Sanfilippo et al. [30] developed a violet intent modeling software (VIM) [30], using a Wikibased expert knowledge management approach and content extraction and content extraction, which describe a framework that implemented a multidisciplinary approach in the emergence of radicalization leading to violet intent based on English language only.

3 Problem Definition

The increase in the emerging trend of hate groups, and their presence online in recent years call for concern. In 2014, a survey conducted by the New York Times with the Police Executive Research Forum reported that rightwing extremism is the primary source of "ideological violence" in America and that Europe is dazed by the rise of far-right extremist groups [31]. From literature review, there is no single study that has investigated the methods of identifying hate speech in multilingual languages, in terms of word order and n-gram concept. With features concept analysis, as against a key word technique, there is need to apply an improved feature model for extraction, and a context-based technique that combines bi-grams to accurately classified intent, hence the computational research problem as stated below:

In this study, hate detection will be modeled for bi-lingual posts. We assume that we are interested in classifying a dataset D of twitter feeds t, represented in Eq. 5.

$$\mathrm{D}^{l} = \mathrm{t_i}\{1| \leq \mathrm{i} \leq \mathrm{n}\} \quad \text{where } l \in \{\mathrm{l_1}, \mathrm{l_2}\}. \tag{5}$$

D_1^l and D_2^l are two distinct documents with two distinct different languages l_1 and l_2.

The first is to preprocess D with n-gram concept, using Topical n-gram (TNG) model to identify a set of topics, C, for each tweet, where $\mathrm{C} = \{\mathrm{c_1}, \mathrm{c_2}, \ldots, \mathrm{c_k}\}$. Then we assign the mixture of topics to n-gram phrases of words to get a set of predefined hate terms, S. Finally, we classify vectored terms. Our goal is to define f which can classify the set of terms to either True or False.

$$f: S \rightarrow \{True, False\} \text{ where } f \text{ is the unknown intent function.}$$

3.1 Proposed Methodology

Here we propose a framework for the model for classification of hate intent of multi-lingual twitter posts. The framework is represented in Fig. 1.

The proposed frame work in Fig. 1, will take two different twitter dataset that is bilingual, t_i^1 and t_j^2. Each dataset will be subjected to pre-processing by cleaning the bilingual tweets of non-textual information and not relevant topics. Next, stop-word removal as well as tokenization, stemming, Part of Speech tagging, and lemmatization, using appropriate lexicons and semantics appropriate for each dataset will be carried out.

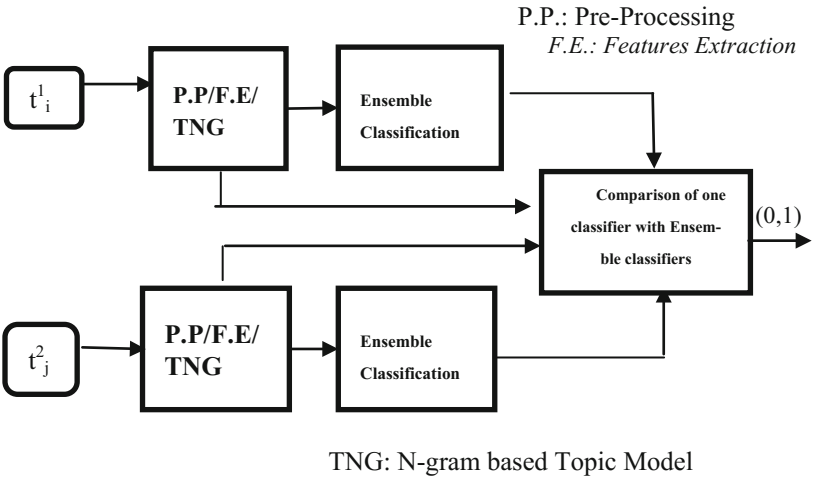

Fig. 1. The structural framework of the proposed solution to the problem statement.

The feature extraction will include Expected Cross Entropy (ECE) and Gibbs Sampling for vectorization, while supervised context-based N-gram topic model (TNG) will be employed for topic modeling. One term keyword spotting based techniques cannot be used to accurately identify the intent of a post, it requires context-based techniques that combines bi-grams to accurately classified intent as this can better defines intent in a post. Feature extraction and topic modeling will be employed for the two datasets separately, so also with the Ensemble Classification using Naïve Bayes, and Support Vector Machine for classification. The performance of each classifier will be assessed using precision, recall and accuracy and performance will also be compared with existing methods.

4 Conclusion

Existing literature have shown that hate and radicalism speech are not tied to only minor communities, but larger communities. Researchers have shown much concern to commercial intent analysis, while very few attentions are shown towards social hate intent classification in multilingual natural language text. While Sentiment serves as a rationale for emotional expressions, hate intent can characterize a person's goal and thus provides additional information about the person itself. One term keyword-based techniques cannot accurately identify the intent of a post. Hence, the focus of this study was to develop a framework for a cascaded ensemble learning based classifier for hate intent classification in multilingual languages using topical n-gram model.

We have proposed a framework for intent identification in bilingual twitter post, using context-based technique. If implemented, the algorithm will help in ensuring national, social, and information security.

References

1. Kröll M, Strohmaier M (2015) Associating intent with sentiment in weblogs. In: International conference on applications of natural language to information systems. Springer, Cham
2. Albright J (2016) The# Election2016 micro-propaganda machine. https://medium.com/@d1gi/the-election2016-micro-propaganda-machine-383449cc1fba#.idanl6i8z. Accessed 15 Jan 2017
3. Lozhnikov N, Derczynski L, Mazzara M (2018) Stance prediction for Russian: data and analysis
4. Shu K et al (2017) Fake news detection on social media: a data mining perspective, 19 (1):22–36
5. Purohit H et al (2015) Intent classification of short-text on social media. In: 2015 IEEE international conference on smart city/SocialCom/SustainCom (SmartCity). IEEE
6. Kröll M, Strohmaier M (2009) Analyzing human intentions in natural language text. In: Proceedings of the fifth international conference on knowledge capture. ACM
7. Mohtarami M et al (2018) Automatic stance detection using end-to-end memory networks
8. Zubiaga A et al (2018) Detection and resolution of rumours in social media: a survey, 51 (2):32
9. Lozhnikov N, Derczynski L, Mazzara M (2018) Stance prediction for russian: data and analysis. arXiv preprint arXiv:1809.01574
10. Dai HK et al (2006) Detecting online commercial intention (OCI). In: Proceedings of the 15th international conference on world wide web. ACM
11. Kirsh D (1990) When is information explicitly represented? Information, language and cognition - the Vancouver studies in cognitive science. UBC Press, pp 340–365
12. Ajzen I (1991) The theory of planned behavior, 50(2):179–211
13. Malle BF, Knobe J (1997) The folk concept of intentionality, 33(2):101–121
14. Sloman SA et al (2012) A causal model of intentionality judgment, 27(2):154–180
15. Melnikov A et al (2018) Towards dynamic interaction-based reputation models. In: 2018 IEEE 32nd international conference on advanced information networking and applications (AINA). IEEE
16. Hollerit B, Kröll M, Strohmaier M (2013) Towards linking buyers and sellers: detecting commercial intent on Twitter. In: Proceedings of the 22nd international conference on world wide web. ACM
17. Benczúr A et al (2007) Web spam detection via commercial intent analysis. In: Proceedings of the 3rd international workshop on adversarial information retrieval on the web. ACM, pp 89–92
18. Lewandowski D, Drechsler J, Von Mach S (2012) Deriving query intents from web search engine queries. J Am Soc Inform Sci Technol 63(9):1773–1788
19. Guo Q, Agichtein E, Clarke CL, Ashkan A (2008) Understanding "abandoned" ads: towards personalized commercial intent inference via mouse movement analysis. Inf Retr Advert IRA 2008:27–30
20. Lewandowski D (2011) The influence of commercial intent of search results on their perceived relevance. In: Proceedings of the 2011 iConference. ACM, pp 452–458
21. Ben-David A, Matamoros-Fernandez A (2016) Hate speech and covert discrimination on social media: monitoring the Facebook pages of extreme-right political parties in Spain. Int J Commun 10:1167–1193
22. Wang X, McCallum A, Wei X: Topical n-grams: phrase and topic discovery, with an application to information retrieval. In: ICDM. IEEE, pp 697–702

23. Chavhan RN (2016) Solutions to detect and analyze online radicalization, 1(4)
24. Agarwal S, Sureka A (2017) Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on Tumblr micro-blogging website
25. Montejo-Ráez A et al (2014) A knowledge-based approach for polarity classification in Twitter, 65(2):414–425
26. Balahur A, Perea-Ortega JM (2015) Sentiment analysis system adaptation for multilingual processing: the case of tweets, 51(4):547–556
27. Montoyo A, MartíNez-Barco P, Balahur A (2012) Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments. Elsevier (2012)
28. Vilares D et al (2017) Supervised sentiment analysis in multilingual environments, 53 (3):595–607
29. Gomes HM et al (2017) A survey on ensemble learning for data stream classification, 50 (2):23
30. Sanfilippo A et al (2009) VIM: a platform for violent intent modeling. In: Social computing and behavioral modeling. Springer, Heidelberg, pp 1–11
31. Ben-David A, Matamoros-Fernandez A (2016) Hate speech and covert discrimination on social media: monitoring the Facebook pages of extreme-right political parties in Spain, 10:1167–1193