Unsupervised and Language Independent Approach to Extremism and Collective Radicalization Understanding

Proposta de Dissertação de Mestrado

Orientador: Sebastião Pais
Co-Orientador: João Paulo Cordeiro

Departamento de Informática @ UBI
2018/2019

1 Context

Each cluster of tweet messages focusing on a bur sty topic may constitute a potential threat. However, the overwhelming majority of clusters are arm less and represent casual, conventional or expressive crowds as well as noisy data (Becker et al., 2011). To identify acting or protest crowds, we propose to understand the typical language usage present in each cluster as well as its network activity. Indeed, ultimately, a crowd is characterized by its dominant emotion, its level of interaction and shared focus.
(Krumm, 2015) showed that specific radicalized language is used within acting and protest crowds. Therefore, we propose that each tweet inside a cluster is classified as radical or non-radical in terms of language use, so that the collective radicalization of a cluster can be measured. As far as we know, there exists no previous work on modeling radicalized language. Radicalization is a process by which an individual or group comes to adopt increasingly extreme political, social, or religious ideals and aspirations . As such, we hypothesize that radicalized language mainly expresses negative emotions (such as anger, fear, or anxiety) with high intensity, following the classification of Plutchik's wheel of emotion (Plutchik, 1980).
There exists a great deal of works aiming at classifying tweets in some intended categories. In all cases, the main problem lays in the fact that only a few training examples (manually labeled by some expert) can be afforded, and so robust classifiers cannot be built. Traditional methods to solve this problem include self-taught, semi-supervised or ensemble learning. In this specific context, we propose a solution based on self-taught learning (Raina et al, 2007).
Self-taught learning aims at defining higher-level feature representations based on unlabeled data that can easily be gathered. Word embeddings (Mikolov et al., 2013) classically propose such high-level representations. Some successful results have recently been obtained in the domain of sentiment tweet classification (Tang et al., 2014), upon which we propose to improve. Although word embeddings have shown great improvements over most NLP tasks, they show some

limitations. As most models exclusively use syntactic contexts (Mikolov et al., 2013), words with different connotations may not be well-separated in the space, and consequently produce error-prone classifiers. Emotional words such as fear and joy are such an example.

2 Goals

This thesis propose to learn weak classifiers based on dictionaries of emotional words (Qadir and Riloff, 2014) to retrieve many roughly classified emotional tweets, in a similar way as (Tang et al., 2014). The continuous high-dimensional space can then be learned by including both syntactic and emotional contexts into a recurrent neural network. However, only highly intensive emotional language may identify threatening crowds and not just emotional language. Intensive language can be seen as the difference between "angry" and "wild". Both words share some common semantics but with different intensity levels. In a well-behaved semantic space, angry and wild should also be separated, which is not the case by only looking at the syntactic and emotional contexts. Therefore, we propose to build word embeddings that consider emotional and intensity contexts at the same time together with syntactic context. For that purpose, weak classifiers for intensity detection will be built based on recent work of Sharma et al. (2015) on language intensity. It is interesting to notice that we will build a continuous semantic space, where both lexical items and named entities are joined. As such, entities such as Adolf Hitler and Winston Churchill may clearly be separated in such a semantic space.
Finally, this thesis will study the introduction of demographically-driven word embeddings. Recently, (Bamman et al., 2014) proposed to develop word embeddings considering the localization of the issuer of the conveyed message. These findings open a great deal of improvements, as tweets can be geo-localized but also include information such as age and gender of the issuer. We deeply believe that localization and age can drastically improve the correct encoding of words in continuous spaces and therefore produce high-performing classifiers for radicalized language. Indeed, radicalism may not be expressed in the same way whether the issuer of a tweet is 18 or 45 years-old. Moreover, geo-localization can consider specific regional language usage. As far as we know, there exist no study on demographically-driven word embeddings. This thesis propose to build word embeddings in a similar way as (Bamman et al., 2014) but including both age and gender as well as geo-localization and all its combinations.

3 Tasks

1. Review the State-of-the-art, write the survey about this problematic;
2. Propose a new Unsupervised and Language Independent Approach to Extremism and Collective Radicalization Understanding;
3. Implementation;
4. Testing and evaluation;
5. The writing of the dissertation.

4 Bibliography

(Bamman et al., 2014) Bamman, B., Dyer, C. and Smith, N.A. (2014). Distributed Representations of Geographically Situated Language. 52nd Annual Meeting of the Association for Computational Linguistics (ACL). pp. 823-834.
(Becker et al., 2011) Becker, H., Naaman, M., & Gravano, L. (2011). Beyond Trending Topics: Real-world Event Identification on Twitter. 5th International AAAI Conference on Weblogs and Social Media (ICWSM). pp. 438-441.
(Krumm, 2015) Krumm, J. S. (2015). The Influence of Social Media on Crowd Behavior and the Operational Environment. Maroon Ebooks.
(Plutchik, 1980) Plutchik, R. (1980). Emotion: Theory, Research, and Experience. New York: Academic.
(Raina et al., 2007) Raina, R., Battle, A., Lee, H., Packer, B. and Ng, A.Y. (2007). Self-taught Learning: Transfer Learning from Unlabeled Data. 24th International Conference on Machine Learning (ICML). pp. 759-766.
(Mikolov et al., 2013) Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems (NIPS). pp. 3111-3119.
(Tang et al., 2014) Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., and Qin, B. (2014). Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. 53rd Annual Meeting of the Association for Computational Linguistics (ACL). pp. 1555-1565.
(Qadir and Riloff, 2014) Qadir, A., and Riloff, E. (2014). Learning Emotion Indicators from Tweets: Hashtags, Hashtag Patterns, and Phrases. Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1203-1209.

5 Contactos

Sebastião Pais (sebastiao@di.ubi.pt) - Gabinete 4.1
João Paulo Cordeiro (jpaulo@di.ubi.pt) - Gabinete 4.3

UBI, Departamento de Informática
Rua Marquês d'Ávila e Bolama
6201-001 Covilhã