SmokPro: Towards Tobacco Product Identification in Social Media Text

Venkata Himakar Yanamandra *, Kartikey Pant *, and Radhika Mamidi

International Institute of Information Technology, Hyderabad
{himakar.yv,kartikey.pant}@research.iiit.ac.in
radhika.mamidi@iiit.ac.in

Abstract. In this work, we explore the fine-grained classification of tweets involving tobacco focused on identifying tobacco products. We release the *SmokPro* dataset, along with an extensible method of labeling the tweets through a comprehensive annotation schema. We then perform benchmarking experiments using state-of-the-art text classification models, exploiting contextual word embeddings and achieve F1 scores as high as 0.971, hence showing the efficacy of the dataset and the suitability of the models for the task.

1 Introduction

Smoking is one of the leading causes of preventable death with tobacco use causing more than 7 million deaths per year worldwide [1]. While cigarette smoking went down among high school students from 2011 to 2019, the number of students using e-cigarettes rose from 3.6 million to 5.4 million [2]. Consequently, 2807 e-cigarette induced lung injury cases were reported in the United States alone, of which, there have been 68 deaths [3]. An e-cigarette search has a high chance of coming across a tilted conversation or an advertisement encouraging e-cigarette use as a socially acceptable practice [10].

It thus becomes essential to monitor and regulate its use through the detection of its mentions in social media text. Even though Twitter is being used as a resource for public health surveillance, very little information is known about tobacco products, especially modern tobacco products [12,13]. There is a wide availability of user-generated content on Twitter, and it is critical to find trends in different tobacco products for further research, monitoring, and regulatory enforcement efforts [9]. The language used in tweets is diverse, idiosyncratic, sprinkled with emojis, and rapidly evolving [3]. In contrast to previous studies, we have taken slang into account as well. The use of variable spellings with non-standard grammar has risen in informal media. Studying this will help us filter ambiguous and sarcastic tweets [4]. Although Pant et al. [13] explored the

[*] The first two authors contributed equally to the work.

[1] https://bit.ly/2WOuQki

[2] https://bit.ly/2UrBnjf

[3] https://bit.ly/3byz6Zf

fine-grained classification of tobacco-related tweets, they did not focus on the type of tobacco product mention, modern or traditional, to be identified. We, therefore, extend their study by annotating their released dataset *SmokEng* on a different dimension.

In this work, we explore tobacco product identification and release the *SmokPro* dataset [4]. We also give the annotation schema used to label the dataset, enabling the dataset to be extensible. We further use existing state-of-the-art methods for text classification including contextual word embeddings to solve the multi-class classification problem effectively.

2 Related Work

There has been significant work towards the identification of tobacco-related social media content. In [12], the authors explored content and sentiment analysis of Tobacco-related Twitter posts. They used basic statistical classifiers for detection with emphasis on emerging products like hookah and e-cigarettes. However, the number of keywords was limited, and they do not consider slang into their dataset collection pipeline. Cole-Lewis et al. worked on content analysis for identifying trends of e-cigarette related tweets[2]. However, they do not give an explicit fine-grained pipeline for the identification of tobacco products. In [8], a supervised predictive model was developed for automatic identification of proponents of e-cigarettes on twitter using a data set of 1000 independently annotated twitter profiles. ”Proponents” of e-cigarettes were defined to be manufacturers, advocates, and users of e-cigarettes who actively promote the product. They also analyzed the behavior of the selected users but did not detect tweets mentioning e-cigarette use. Fine-grained classification of tobacco-related tweets incorporating slang keywords was explored in [13]. However, they do not differentiate between different types of tobacco products. An analysis of marketing trends of e-cigarettes between 2008 and 2013 was studied in [9]. The authors collected e-cigarette tweets using keywords that were classified into advertising and non-advertisement classes and further sub-classes. However, they only consider e-cigarette tweets and thus do not enable identification of the type of tobacco product.

3 Dataset Creation

3.1 Preliminaries

We use the *clean* version of *SmokEng* dataset [13], which consists of 2116 tobacco-related tweets classified into five distinct classes. The authors collected $7,236,442$ tweets between $1st$ October 2018 to $7th$ October 2018, which represents 1% of the entire twitter feed. They then extract tobacco-related tweets using an

[4] `https://github.com/kartikeypant/smokpro-tobacco-product-classification`

exhaustive list of keywords considering colloquial slang. The authors claim to avoid potential bias based on the day by taking the dataset for a full week. They then prune out non-English tweets and tweets from users with less than 100 followers in an attempt to weed out spam and bot behavior. The data was then manually annotated on the following five classes: ambiguous mentions, personal or anecdotal mention of tobacco use, advisory mention of tobacco products, advertisements of tobacco products, and mention of non-tobacco drugs.

3.2 Dataset Annotation

We annotate the data based on five categories for the task of tobacco product identification. These categories were motivated by the general perception of tobacco, e-cigarettes, and non-drug related tweets.

To detect whether a tweet is associated with a tobacco product(traditional or modern), we formulate the following guidelines:-

- Accounts of either personal use of the tobacco products, or provide instances of the use of the products by themselves or others
- Mention statistics of tobacco consumption
- Referring to associated health risks
- Social campaigns condemning the usage of the tobacco
- Endorsing or targeting the sale of the tobacco products and analogous products or services

We then label each tweet as either of the following classes:

1. **Traditional Tobacco Product Mention**: Tweets containing a tobacco product mention according to the above guidelines related to a traditional tobacco product. We consider *cigarette, hookah, pipe, cigar, bidis, cigarillo, shisha,* and *baccy* as the traditional tobacco products.
2. **Modern Tobacco Product Mention**: Tweets containing a tobacco product mention according to the above guidelines related to a modern tobacco product. We consider *e-cigarette, e-juice, e-hookahs, e-liquid, mods, vape pens, vapes, tank systems,* and *electronic nicotine delivery systems (ENDS)* as the modern tobacco products.
3. **Generic Mention of Smoking**: Tweets portraying the aforementioned definitions of tobacco usage without referring to any kind of tobacco products be it traditional or modern or other drugs.
4. **Narcotics & Other Drug Mentions**: Tweets denoting usage, purchase, and information about narcotics and drugs other than traditional and modern tobacco products[5].
5. **Ambivalent or Unclear Mentions**: This category of tweets contain tweets containing information unrelated to tobacco or any other drug, or about ambiguity in the intent of tweet, such as sarcasm.

[5] https://www.incb.org/incb/en/narcotic-drugs/index.html

Category	Example	Count
Ambivalent Mentions	”$MENTION$ $MENTION$ What have you been smoking”	333
Traditional Tobacco Mentions	”Trying to sell cigars to a non-smoker is a waste of time. Identify your target audience and create your message spe... URL”	555
Narcotics Mentions	”$MENTION$ Making my money and smoking my weed”	393
Modern Tobacco Mentions	”my vape stopped working Friday so I took it back and all he would do is send it back to the company!!! I'm furious.”	293
General Mentions	”Smoking alone is boring lol”	542

Table 1. Class-wise distribution of tweets, with an example per class.

3.3 Inter-annotator agreement

We calculate the Inter-Annotator Agreement (IAA) to assess the quality of annotation for the fine-grained classification between the two annotation sets of $2,116$ tobacco-product related tweets using Cohen's Kappa coefficient [6]. The Kappa score of 0.861 indicates that the high quality & usefulness of the schema.

4 Methodology

In this section, we describe the classifiers designed for the task of fine-grained classification. We employ widely used contextual word embedding models like BERT and RoBERTa to obtain state-of-the-art performance for this task of multi-class text classification. We also use carefully tuned FastText to serve as a baseline in our comparison.

We use the following models for the experiments:-

1. **FastText**[7]: *Fasttext* classifier uses embeddings for each character n-gram which helps in improving the task on newer unseen data as it captures information about local word ordering. We train the model in an end to end fashion by reducing the cross-entropy loss over the predictions using an SGD optimizer [1].
2. **BERT**[5]: *BERT* is contextualized word representation based on bidirectional transformers. that leverages context from both left and right representations in each layer. The model can be trained in a simpler, yet efficient manner without having to make major architectural changes. BERT is classified into two types, based on the casing characteristic of the input: *Uncased* $BERT_{Large}$ and *Cased* $BERT_{Large}$ Both models are based in $BERT_{Large}$ which uses a 24-layered transformer with $340M$ parameters. We finetune both models on the training dataset and evaluate the finetuned models on the test dataset.
3. **RoBERTa**[11]: *RoBERTa* is a replication study of *BERT*, trained on a much larger dataset of over $160GB$ training dataset as compared to $16GB$ training dataset used for *BERT*. While *BERT* uses character-level encodings for training, *RoBERTa* makes use of larger byte-pair encoding(BPE) vocabulary that helps to achieve better performance on various tasks. Finally, compared

to *BERT*, it removes the next sequence prediction objective from the training procedure. We use its large variant, $RoBERTa_{large}$, and finetune it for the task.

5 Experiments and Results

In this section, we describe the full fine-grained classification experiment towards detection of tobacco product mentions in the *SmokPro* dataset. The experiment is designed to show the efficacy of existing state-of-the-art models for text classification tasks in our dataset.

For both *BERT*-based and *RoBERTa* models, we use a learning rate of $2 * 10^{-5}$, a maximum sequence length of 50, and a weight decay of 0.01 while finetuning the model. We use the recently released automatic hyperparameter optimization technique for optimizing *FastText*'s hyperparameters using the validation set.

The experiment conducted in the study entails classification of tweets into the five classes as defined above. We perform all our experiments using a 72-20-8 train-test-validation split of the dataset. Table 2 illustrates the results of the experiment, using the following four metrics: Accuracy, and weighted F1 score. We observe $CasedBERT_{Large}$ to outperform all models obtaining an accuracy of 97.10% and F1 score of 0.971 in the experiments. Further, $RoBERTa_{Large}$ performs competitively obtaining an accuracy of 96.04% and F1 score of 0.960. As a non contextual word embedding based models, FastText performs fairly well obtaining 80.18% and F1 score of 0.801.

Methods	**Accuracy**	**F1 Score**
FastText	80.18%	0.801
Uncased $BERT_{Large}$	92.84%	0.928
Cased BERT_{Large}	**97.10%**	**0.971**
$RoBERTa_{Large}$	96.04%	0.960

Table 2. Experimental results for the five-class tobacco product identification task.

6 Conclusion and Future Work

In this work, we explored identification of tobacco product mention in a given tweet. We propose an extensible data annotation schema for the task. We also release the *SmokPro* dataset, a 2116 tweet dataset manually annotated into five classes as defined by the schema. We then perform experiments using existing state-of-the-art for text classification models in the given dataset and obtain F1 scores as high as 0.971. The effective predictive performance for the task paves way for future work on disease surveillance, personal health mention detection and aspect-based sentiment analysis of social media text pertaining to tobacco products.

References

1. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2016)
2. Cole-Lewis, H., Pugatch, J., Sanders, A., Varghese, A., Posada, S., Yun, C., Schwarz, M., Augustson, E.: Social listening: A content analysis of e-cigarette discussions on twitter. Journal of Medical Internet Research **17**, e243 (10 2015). https://doi.org/10.2196/jmir.4969
3. Collier, N., Nguyen, S., Nguyen, N.: Omg u got flu? analysis of shared health messages for bio-surveillance. Journal of Biomedical Semantics **2** (10 2011). https://doi.org/10.1186/2041-1480-2-S5-S9
4. Conway, M., Hu, M., Chapman, W.: Recent advances in using natural language processing to address public health research questions using social media and consumergenerated data. Yearbook of Medical Informatics **28**, 208–217 (08 2019). https://doi.org/10.1055/s-0039-1677918
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019), https://aclweb.org/anthology/papers/N/N19/N19-1423/
6. Fleiss, J.L., Cohen, J.: The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and Psychological Measurement **33**(3), 613–619 (1973). https://doi.org/10.1177/001316447303300309, https://doi.org/10.1177/001316447303300309
7. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. In: EACL (2016)
8. Kavuluru, R., Sabbir, A.: Toward automated e-cigarette surveillance: Spotting e-cigarette proponents on twitter. Journal of Biomedical Informatics **61** (03 2016). https://doi.org/10.1016/j.jbi.2016.03.006
9. Kim, A., Hopper, T., Simpson, S., Nonnemaker, J., Lieberman, A.A., Hansen, H., Guillory, J., Porter, L.: Using twitter data to gain insights into e-cigarette marketing and locations of use: An infoveillance study. Journal of Medical Internet Research (06 2015). https://doi.org/10.2196/jmir.4466
10. Lazard, A., Saffer, A., Wilcox, G., Chung, A., Mackert, M., Bernhardt, J.: E-cigarette social media messages: A text mining analysis of marketing and consumer conversations on twitter. JMIR Public Health and Surveillance **2**, e171 (12 2016). https://doi.org/10.2196/publichealth.6551
11. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach (07 2019)
12. Myslín, M., Zhu, S.H., Chapman, W., Conway, M.: Using twitter to examine smoking behavior and perceptions of emerging tobacco products. Journal of medical Internet research **15**, e174 (08 2013). https://doi.org/10.2196/jmir.2534
13. Pant, K., Yanamandra, V.H., Debnath, A., Mamidi, R.: Smokeng: Towards fine-grained classification of tobacco-related social media text. In: W-NUT@EMNLP (2019)