1-2020

Using Data Analytics to Filter Insincere Posts from Online Social Networks. A case study: Quora Insincere Questions

Mohammad A. Al-Ramahi
Texas A&M University-San Antonio, mrahman1@tamusa.edu

Izzat Alsmadi
Texas A&M University-San Antonio, ialsmadi@tamusa.edu

Using Data Analytics to Filter Insincere Posts from Online Social Networks A case study: Quora Insincere Questions

Mohammad Al-Ramahi
Texas A&M, San Antonio
mrahman1@tamusa.edu

Izzat Alsmadi
Texas A&M, San Antonio
izzat.alsmadi@tamusa.edu

Abstract

The internet in general and Online Social Networks (OSNs) in particular continue to play a significant role in our life where information is massively uploaded and exchanged. With such high importance and attention, abuses of such media of communication for different purposes are common. Driven by goals such as marketing and financial gains, some users use OSNs to post their misleading or insincere content.

In this context, we utilized a real-world dataset posted by Quora in Kaggle.com to evaluate different mechanisms and algorithms to filter insincere and spam contents. We evaluated different preprocessing and analysis models. Moreover, we analyzed the cognitive efforts users made in writing their posts and whether that can improve the prediction accuracy. We reported the best models in terms of insincerity prediction accuracy.

1. Introduction

In Online Social Networks (OSNs), the content is uncontrolled; users can post, in most cases, in free-form texts; just about anything, they want to say. They can also post information that is entirely fake or insincere. Websites still lack the mechanisms and abilities to check content validity and enforce that; for example, the content could be fake or inaccurate.

Information credibility is a serious problem on the internet. For instance, many references indicated that online products might include fake reviews that are artificially posted to deceive readers. Such reviews seek to either promote products by giving extreme positive reviews (i.e., hyper spam) or damage the reputation of products by providing extreme negative reviews (i.e., defaming spam) [1]. This type of manipulated fake reviews can be particularly harmful in three situations; when (1) they recommend a low-quality product that most other reviewers disagree with, (2) slander a right quality product that most other reviewers like, or (3) incorrectly praise/defame an average quality product [1]. In addition to fake product reviews, users can write posts with fake news and incorrect information as if they are facts or accurate. Such information may get famous and be more visible to search engines than more precise information (i.e., in the same subject or context). For example, a student who is trying to search the internet about a city, an event or a public figure, may hit one of the popular, incorrect articles and use it as if it's the primary, correct information source. In other words, as search engines rank by popularity and not by information accuracy or credibility, using the internet as the primary source of information can cause many problems.

In the context of information credibility, there are three main entities to evaluate: the website containing the post, post author or writer, and the post content. Those three entities depend on each other. For example, a credible website only allows trustworthy authors or contents or have some mechanisms to filter untrustworthy authors and contents. Similarly, trustworthy authors usually post trustworthy information on trustworthy websites.

Should websites be allowed to censure or discipline insincere comments that are harmless? Websites have different conflicting reasons to censor such behaviors or not. They need to balance expanding their audience, focus on quantity, and provide validated content to their loyal users, quality content. Websites that try to deal with such a problem (i.e., information credibility) will face different challenges. Those websites do not want to be seen as "controlling what should be posted," opposing freedom of speech and not allowing their users to express their thoughts or opinions.

On the other hand, the mechanisms to *automatically* detect that a newly created response is incredible are immature and may trigger many false positives or negatives. As an alternative, manual, or human detection and elimination of incredible

URI: https://hdl.handle.net/10125/64046
978-0-9981331-3-3

content require significant time and effort. Certain fact-checking websites such as snopes.com, which are more of claims or fake news assessment website rather than fake reviews' assessment website, dedicate human experts to assess claims and content credibility.

Quora, just like many other OSNs, has credibility issues. Quora is “a platform that empowers people to learn from each other. On Quora, people can ask questions and connect with others who contribute unique insights and quality answers”. Even in comparing Quora with Wikipedia, which has its known credibility issues, many see Wikipedia as containing information/facts, whereas Quora has “opinions." So how can you judge the credibility of opinions?

Nonetheless, it should be mentioned that our focus, as well as the center of Kaggle Quora competition, is not on the credibility of posts, but rather “insincerity” of the question posted. Not only the answers can be insincere, but also the questions as well, especially if users post trivial questions where their goal is only to gain some visibility or popularity. According to the Kaggle website, “an insincere question is a question intended to make a statement rather than look for helpful answers." On Quora, the purpose of Questions is to solicit Answers, not to make statements or advocate viewpoints. Thus, in Quora, it is essential to understand what belongs to the question and what should be in the Answer. Having a Question cited as "Insincere" generally means users put something in the question that should only be in an Answer, like personal views/opinions. A key challenge Quora encounters is to get rid of insincere questions so they can keep their platform a place where people can feel safe sharing their knowledge with others. In this research, we aim to leverage data analytics to predict if a question is sincere or not. Data analytics have been demonstrated as useful tools to analyze user-generated contents in OSNs [e.g., 2, 3-8]. Data analytics can help develop scalable models to detect insincere and misleading content. To this end, we used a unique real-world data set obtained from Quora Website. Specifically, our goals and objectives are:

1) Explore the role of text preprocessing and feature representation in detecting insincere content in online social media.
2) Examine the performance of different supervised machine learning algorithms (e.g., decision tree, linear SVC, logistic regression, and random forest) in detecting insincere contents using diverse data representation.
3) Analyze the cognitive efforts users spend in writing their posts and the role of that in detecting insincere content.

The rest of the paper is organized as the following: Section 2 summarizes a selection of relevant research contributions, and section 3 presents Quora data analysis and our experiment framework. In section 4, we discuss the results on Quora dataset, and paper is concluded in section 5.

2. Literature review

2.1. Information credibility

Information credibility is rarely assessed on the internet due to several reasons, including the lack of quality control mechanisms [9-13]. Credibility can be associated with correctness, truth, or facts. However, much of the content in OSNs convey opinions where there is no reference to correctness. Users in OSNs talk about news events, celebrities, politics, events, fashions, etc.

Many authors looked into cues for deception in OSN posts [14-17]. In OSNs, cues of deceptions that are available for face to face communications (e.g., eye contact, gaze aversion, shrugs, amplitude, etc.) are not applicable [15]. Authors in [15] described a new list of cues that can be used in OSNs deception. Those include sentence length, sentence complexity, sentiment, text informality, emoticon usage, etc. In one finding, they indicated that deceivers usually use short sentences.

Appling et al. [16] described different types of deception strategies, including Falsification, exaggeration, omission, and misleading. Deceptions can also be categorized based on strategies and models and also based on intent to deceive [17].

2.2. Fake reviews

In this section, we will cover a subset of research papers tackled the issue of fake reviews. Fake reviews can be as a result of actual or fake sales. In other words, vendors may seek artificial reviewers to both buy their products and review them, or they may give them incentives to write artificial reviews. On the other hand, vendors may try to inject negative reviews on their rivals.

One issue discussed in fake reviews is the cases of “duplicate or repeated reviews." A significant approach in literature focused on detecting duplicate reviews as the primary indicator for online spam reviews. This approach assumes that such types of reviews are likely to be reposted repeatedly by

spammers. Jindal and Liu [18, 19] used duplicate reviews as positive training data set to build a logistic regression model to detect non-duplicate spam reviews with similar characteristics. To be able to improve the detection accuracy, meta-features about reviews and reviewers should be included. The model is tested against outlier reviews (i.e., reviews with high rating deviation from the average product rating) to check whether it can predict non-duplicate reviews.

In another study, Lau, et al. [20] built a model based on language model probability and "semantic overlapping" to detect semantically similar reviews. To evaluate their model, the authors picked up those reviews with high Cosine similarity as the untruthful candidate set. Then, two experienced annotators were appointed to review the candidate spam set. Approaches that heavily rely on text similarity are only appropriate for certain types of spamming activities when spammers post duplicated or semantically similar reviews on similar or different products.

Instead of using duplicate reviews as evaluation data set, Ott, et al. [21] released hotel reviews data set, which contains 400 truthful reviews obtained from www.tripadviser.com and 400 deceptive positive hotel reviews gathered using Amazon Mechanical Turk (AMT). Based on the data set, authors reported that N-gram-based approach (i.e., N-gram model) is better to detect fake reviews with an accuracy of 90% compared to the two other approaches: genre identification and psycholinguistic deception. However, the words identified by authors as spam indicators are quite typical and thus may appear in any truthful reviews. Feng, et al. [22] extend Ott, et al. [21]'s work by incorporating deep syntax patterns derived from Probabilistic Context-Free Grammar (PCFG) parse trees (i.e., N-Gram + SYN model). They obtained better accuracy on the same data set used by cited authors (91.2%). Feng and Hirst [23] enhanced Feng, et al. [22]'s work by adding profile alignment compatibility features (i.e., C+N-Gram+SYN). These features represent the degree to which aspects mentioned in a review with their descriptions are compatible with those mentioned in the object profile built from all truthful reviews on the object. The results indicated a significant improvement in the performance of identifying deceptive reviews.

Spammers rating behaviors are examined by Lim, et al. [24]. They proposed an aggregated scoring scheme based on four practices to rank reviewers according to their spamming actions. The results indicated that posting multiple similar reviews by a reviewer on either the same products or on products with common attributes such as related brands/products are powerful indicators of spammer behaviors. The study assumes that spammers post multiple similar reviews with the same user identification ID. However, as spammers often adopt obfuscation strategies by changing their user identification when they write several reviews, their behavior would not be detected. Jindal, et al. [25] treat detecting spammer reviewers' problem via formulating their unusual patterns in the data set as finding unexpected rules and rule groups. Such rules associate attributes of the reviews such as reviewer-id, product-id, and brand-id with a particular rating class which can be positive, negative, or neutral. However, the study did not consider that the same reviewer may post several similar reviews but with different user-identifications.

Many other papers cover the subject of "fake reviews" from different aspects, (e.g., Mukherjee, et al. [26], Lappas [27], Malbon [28], and Li, et al. [29]). The problem of fake reviews can be at a large scale orchestrated by groups rather than individuals, Mukherjee, et al. [26]. In comparison with spam detection techniques, fake reviews detection techniques face similar challenges of possible false positive and negative cases. Paper indicated that group-based detection techniques could utilize metrics that measure the level of orchestration in reviews in terms of content agreement or nature, group size and also in terms of the time of occurrence of the "similar" fake reviews. The probability of fake reviews being detected increases with the volume of injected reviews and the ability to detect specific patterns in those reviews.

Lappas [27] focuses on identifying fake reviews and evaluates the impact and authenticity of three factors in reviews: stealth, coherence, and readability. The author regards fake reviews as a form of a malicious attack on reputations. He provides an attacker's perspective on creating authentic-looking and impactful reviews. The paper also showed that some creators of fake reviews adopted approaches to minimize the volumes of fake reviews per product or vendor to avoid detection.

Focusing on investigating methods to handle fake reviews, Malbon [28] discussed the need to take fake reviews as a severe problem. The behavior is shown to be adopted by individuals as well as companies, manufacturers, and/or retailers. In their attempts to influence customers' decision to buy their products, sellers may get attempted to commit some form of fake reviews. While laws and regulations exist to prevent the creation of fake reviews and any other similar deception methods, non-the less, the process to detect such behaviors is not trivial.

Li, et al. [29] construct a user-IP-review graph to detect reviews that are written by the same users and from the same IPs. Authors also studied patterns of posting rates as a method to detect fake reviews. They also utilized sentimental analysis and the trends in the polarity of reviews (i.e., positive or negative reviews) as a method to detect fake reviews.

3. Data and experiments' framework

In this section, we explain our data and experiment framework towards the goal of identifying methods to detect insincere contents in Quora dataset.

3.1. Data

The data used in this study was obtained from the Quora Website (https://www.kaggle.com/c/quora-insincere-questions-classification/data). Each record in the data includes the question that was asked, and whether it was identified as insincere (target = 1) or not (target = 0).

3.2. Experiments' framework

Our goal is to explore the role of text preprocessing and feature selection/representation in detecting insincere content on social media. To this end, we conducted two experiments, as shown in Figure 1. We used Python as a data analytic language/tool to implement both experiments. We used one data set of Quora questions randomly split into two smaller data sets; the first one was used in the first experiment that contains 60,768 questions (30,581 insincere and 30,187 sincere) (data set 1), and the other one was used in the second experiment with 15,004 questions (7,825 insincere and 7,179 sincere) (data set 2). The reason we divided the dataset into two parts is that in the second experiment, we used N-gram representation which generates much more significant feature space than a traditional bag of words representation that is used in experiment 1. Therefore, and due to our memory size limitation, in experiment 2, we used a smaller data set to reduce the number of features (i.e., N-grams) generated.

Both experiments consist of four key stages: (1) questions preprocessing (stop words removal and stemming), (2) feature representation and feature selection, (3) classification process and (4) performance evaluation.

3.2.1. Data preprocessing. To examine whether stemming improves the prediction of insincere posts, we evaluated two different preprocessing techniques, as shown in Figure 1 (see experiment 1). First, stop words are removed, then stemming is applied. Stemming is the process of converting words that are in their inflected forms (e.g., plural nouns and past-tense verbs) to their original forms. Second, we just removed the stop words (i.e., no stemming is performed).

3.2.2. Data representation. Questions were then represented using different features. For example, we used the bag of words (i.e., unigrams) as features in experiment 1. In experiment 2, we added bi-grams and tri-grams to the uni-grams to compare the performance against the unigram features only in the first experiment. For example, 'Quora,' 'Quora questions,' and 'Quora insincere questions' are examples of unigram, bigram, and trigram, respectively. After that, different feature matrices were constructed for each one of the datasets based on three different types of feature weighting methods: Term Presence (TP), Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF). In the TP matrix, the *(i, j)*-th entry is the weight of feature I in question *j* (i.e., one, if the feature exists and 0 otherwise). In the TF matrix, the weight is the frequency of feature I in question *j*. The formula used for TF-IDF is: TF + (TF * IDF), instead of TF * IDF.

Specifically, TF-IDF weight of a feature *i* in a document *j* is:

$$TF_{i,j} + (TF_{i,j} * log(N/DF)) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

Where $TF_{i,j}$ is the frequency of the feature I in the question j and N indicates the number of questions in the corpus. DF is the number of questions that contain feature i. The effect of this is that features with zero IDF, i.e., that occur in all questions of a training set will not be entirely ignored. TF is normalized using the sum of all TFs in the question or the post.

3.2.3. Feature selection. One problem with representing the questions as vectors of uni-grams (i.e., the bag of words) is a large number of generated features. The problem will be even worse when including bi-grams and tri-grams as we did in experiment 2. Such a vast number of features can potentially cause model or results' overfitting. We, therefore, performed feature selection using the commonly used Chi-square (X^2) method. The Chi-square method evaluates features individually by measuring their Chi-square statistics concerning the classes of the target variable (i.e., insincere or

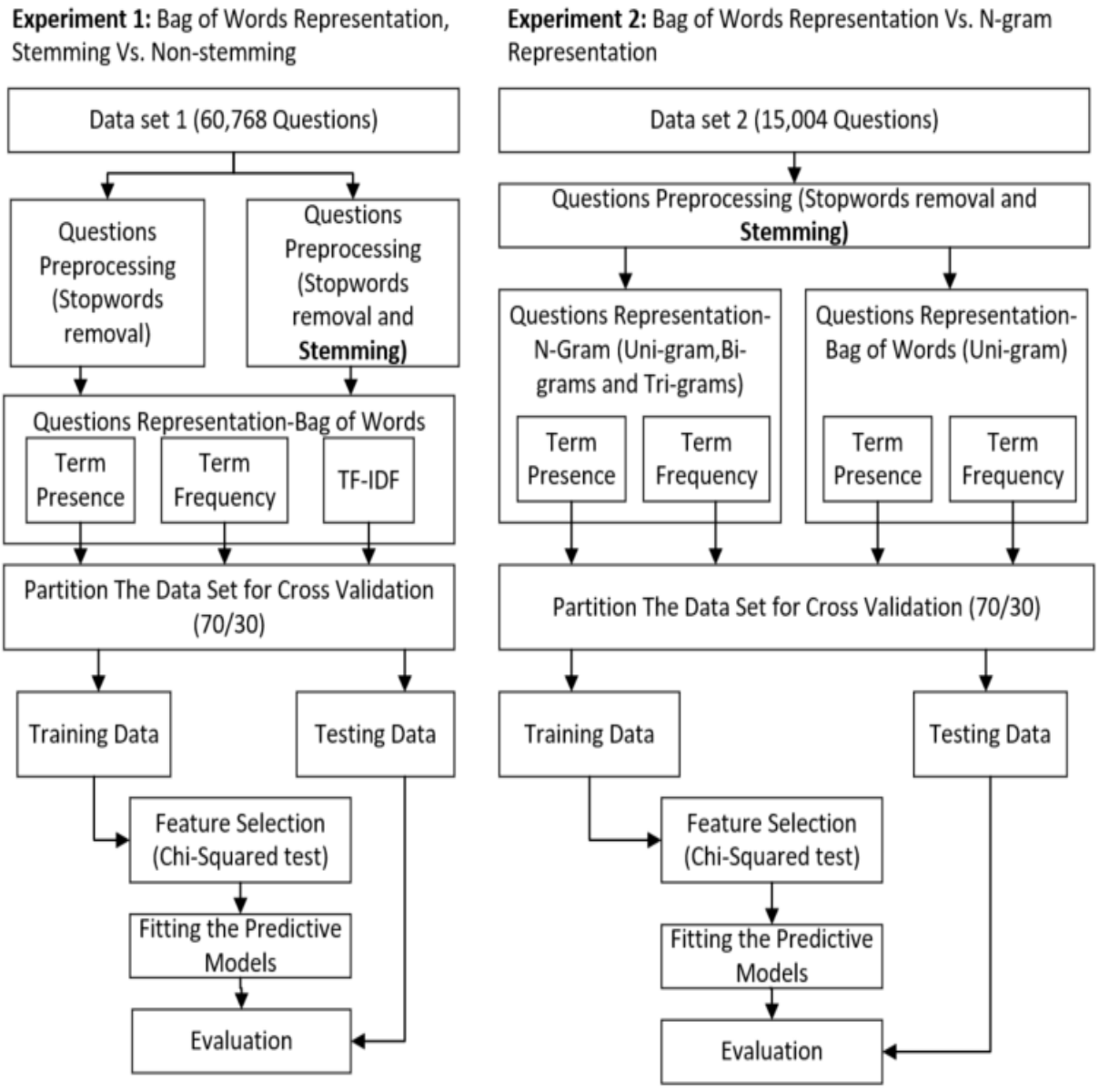

Figure 1: Our Experiments' Framework

sincere). As a result, we only selected the features that have a Chi-square test score that is statistically significant at the 0.05 level (i.e., p-value <0.05). As a result, the number of features was significantly reduced. Since feature selection must be performed using only the training data, we randomly split our data set into 70% training and 30% testing partitions. The training dataset is used for feature selection, and test data is used for evaluation.

3.2.4. Classification process and performance evaluation. After constructing the matrices mentioned above, we evaluated different classifiers on each one of the feature matrices resulting from each data preprocessing and representation. Classifiers used in our experiments include Decision tree, linear SVC, logistic regression, and random forest. We choose these standard and primitive data mining models with their default parameters to establish a few baselines models. To evaluate the predictive power of the selected features, we chose four evaluation metrics, precision, recall, accuracy,

and F1 score. The precision metric evaluates the prediction accuracy by dividing the number of correctly predicted positive samples (TP) on the total number of both TP and FP (those that are mistakenly classified as positive). Note that the drawback of the precision is that it does not account for those who are incorrectly classified as negative samples (i.e., FN).

$$Precision = TP / (TP+FP) \quad \ldots\ldots (2)$$

On the other hand, the recall metric evaluates the prediction accuracy by dividing the number of TP on the total number of both TP, and those are incorrectly classified as negative (FN).

$$Recall = TP / (TP+FN) \quad \ldots\ldots (3)$$

The accuracy metric measures the percentage of those correctly classified as positive or negative examples.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (4)$$

The last metric is F1 score. F1 score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

$$F1\ Score = 2*(Recall * Precision) / (Recall + Precision) \quad \ldots\ldots (5)$$

3.2.5. Cognitive effort analysis. To be able to explore whether sincere and insincere questions are different in length, the length of a question in sentences, words, and characters were added. These features were chosen since they measure the cognitive effort that a user invests in writing a question [30]. Users are expected to put more cognitive efforts in writing sincere questions in comparison with the insincere ones.

4. Results and discussion

Table 1 and Table 2 show the results of the first and second experiments, respectively.

Table 1: Experiment 1 results, Bag of Words Representation, Stemming Vs. Non-stemming

Stemming				
Use term presence instead of term frequency				
Algorithm	F1	Accuracy	Precision	Recall
LinearSVC	0.8628	0.8634	0.8482	0.8779
LogisticRegression	**0.8656**	**0.8664**	**0.8496**	**0.8822**
DecisionTree	0.7955	0.8015	0.7628	0.8312
RandomForest	0.8185	0.8210	0.7973	0.8408
Use term frequency instead of term presence				
Algorithm	F1	Accuracy	Precision	Recall
LinearSVC	0.8623	0.8631	0.8465	0.8786
LogisticRegression	**0.8655**	**0.8666**	**0.8482**	**0.8836**
DecisionTree	0.7917	0.7986	0.7559	0.8312
RandomForest	0.8189	0.8221	0.7947	0.8446
TF-IDF				
Algorithm	F1	Accuracy	Precision	Recall
LinearSVC	0.8187	0.8254	0.7786	0.8631
LogisticRegression	0.8091	0.8168	0.7667	0.8563
DecisionTree	0.7937	0.7948	0.7797	0.8082
RandomForest	0.8109	0.8132	0.7915	0.8313
Non-Stemming				
Use term presence instead of term frequency				
Algorithm	F1	Accuracy	Precision	Recall
LinearSVC	0.8622	0.8635	0.8433	0.8819
LogisticRegression	0.8675	0.8691	0.8465	0.8896
DecisionTree	0.7876	0.7976	0.7411	0.8402
RandomForest	0.8132	0.8190	0.7783	0.8515
Use term frequency instead of term presence				
Algorithm	F1	Accuracy	Precision	Recall
LinearSVC	0.8597	0.8611	0.8403	0.8800
LogisticRegression	0.8659	0.8679	0.8422	0.8909
DecisionTree	0.7855	0.7970	0.7343	0.8444
RandomForest	0.8089	0.8152	0.7727	0.8486
TF-IDF				
Algorithm	F1	Accuracy	Precision	Recall
LinearSVC	0.8100	0.8205	0.7561	0.8723
LogisticRegression	0.7941	0.8070	0.7353	0.8632
DecisionTree	0.7896	0.7947	0.7612	0.8204
RandomForest	0.8032	0.8095	0.7678	0.8420

Experiment 1 results show that stemming process achieves approximately similar performance over non-stemming (for example, using TP feature representation, F1: 0.8656 vs. 0.8675, Accuracy: 0.8664 vs. 0.8691, Precision: 0.8496 vs. 0.8465, Recall: 0.8822 vs. 0.8896) with very slightly better performance for non-stemming especially in terms of recall. As a result, we can see that stemming is not an essential preprocessing step in predicting insincere questions. Experiment 1 results also report a significant performance for TP and TF data representation against TF-IDF. Experiment 2 results reveal that including bi-grams and tri-grams features will not enhance the performance of the classifiers. Finally, Logistic Regression achieved better performance against other classifiers followed by linear SVC.

Table 2: Experiment 2 results, Bag of Words Representation Vs. N-gram Representation

Bag of Words Representation				
Use term presence instead of term frequency				
Algorithm	F1	Accuracy	Precision	Recall
LinearSVC	0.8471	0.8438	0.8229	0.8727
LogisticRegression	**0.8561**	**0.8538**	**0.8276**	**0.8868**
DecisionTree	0.7636	0.7694	0.7084	0.8281
RandomForest	0.7908	0.7930	0.7447	0.8431
Use term frequency instead of term presence				
Algorithm	F1	Accuracy	Precision	Recall
LinearSVC	0.8413	0.8387	0.8132	0.8714
LogisticRegression	**0.8497**	**0.8476**	**0.8195**	**0.8822**
DecisionTree	0.7630	0.7677	0.7113	0.8226
RandomForest	0.7979	0.7970	0.7625	0.8367
N-gram Representation (Unigram, bigrams, and trigrams)				
Use term presence instead of term frequency				
Algorithm	F1	Accuracy	Precision	Recall
LinearSVC	0.8448	0.8416	0.8199	0.8711
LogisticRegression	**0.8531**	**0.8510**	**0.8233**	**0.8851**
DecisionTree	0.7681	0.7737	0.7134	0.8319
RandomForest	0.7959	0.7965	0.7549	0.841
Use term presence instead of term frequency				
Algorithm	F1	Accuracy	Precision	Recall
LinearSVC	0.8382	0.8356	0.8102	0.8682
LogisticRegression	**0.8489**	**0.8470**	**0.8178**	**0.8824**
DecisionTree	0.7711	0.7754	0.7198	0.8303
RandomForest	0.8103	0.8085	0.7781	0.8453

Table 3 shows the results of adding the length meta-features (i.e., the length of a question in sentences, words, and characters) to the prediction model (i.e., logistic regression). Results revealed that adding these features did not improve the model prediction results. Therefore, the length of the questions posted is not significantly correlated with the target. This indicates that sincere or insincere questions cannot be used as significant features to distinguish sincere from insincere questions.

Table 3: Length meta-features (cognitive effort analysis)

Bag of words-term presence representation				
Prediction results WITHOUT length meta-features				
Algorithm	F1	Accuracy	Precision	Recall
LogisticRegression	0.8481	0.8478	0.8085	0.8918
Prediction results WITH length meta-features				
Algorithm	F1	Accuracy	Precision	Recall
LogisticRegression	0.8451	0.8449	0.8051	0.8893

5. Conclusion

In this paper, we investigated the problem of detecting Quora insincere questions as a case study of detecting insincere contents in online social media. We tried a combination of different preprocessing and feature representation methods in addition to using the chi-squared method to remove irrelevant features. We have reported extensive results showing that (1) the appropriate feature representation and filtering in addition to (2) the usage of appropriate classifiers can significantly enhance the accuracy of the prediction process. Specifically, our model showed that the bag-of-words representation with Term Presence (TP) or Term Frequency (TF) weighting scheme is an appropriate representation or model for Quora data. Additionally, results reported that stemming is not an essential preprocessing step in predicting insincere posts.

Further, our analysis showed that logistic regression is an appropriate predictive model to identify insincere questions. Moreover, we added cognitive efforts related features to the model in trying to improve the detection accuracy. However, we noticed that these features are not correlated with the class and hence are not good predictors. Therefore, we conclude that insincere users spend almost the same cognitive efforts in writing insincere questions similar to those who write sincere ones. To best of our knowledge, the techniques reported in our analytical framework were applied for the first time in this context (i.e., detecting Quora insincere questions).

In our future work, we will evaluate a deeper set of features like typos and their impact on models' prediction. The objective is to establish more advanced models and compare them against the baseline models in this paper. Candidate features are those related to the readability and quality of questions posted. For examples, we plan to evaluate the number of spelling errors in the question and the Automated Readability Index (ARI) for the reviews. It would be interesting to see what attributes are the most helpful in predicting an insincere question. Further, the generalizability of the findings will be examined against other OSN platforms such as fact-checking websites. The goal is to explore if the best models for Quora dataset will also be the most accurate when applied to a different OSN.

6. References

[1] A. A. Sheibani, "Opinion mining and opinion spam: A literature review focusing on product reviews," in *6th International Symposium on Telecommunications (IST)*, 2012, pp. 1109-1113: IEEE.

[2] J. A. Chevalier and D. Mayzlin, "The effect of word of mouth on sales: Online book reviews," *Journal of marketing research*, vol. 43, no. 3, pp. 345-354, 2006.

[3] K.-Y. Goh, C.-S. Heng, and Z. Lin, "Social media brand community and consumer behavior: Quantifying the relative impact of user-and marketer-generated content," *Information Systems Research*, vol. 24, no. 1, pp. 88-107, 2013.

[4] Y. Guo, S. J. Barnes, and Q. Jia, "Mining meaning from online ratings and reviews: Tourist satisfaction analysis using latent dirichlet allocation," *Tourism Management*, vol. 59, pp. 467-483, 2017.

[5] M. A. Al-Ramahi, J. Liu, and O. F. El-Gayar, "Discovering Design Principles for Health Behavioral Change Support Systems: A Text Mining Approach," *ACM Transactions on Management Information Systems (TMIS)*, vol. 8, no. 2-3, p. 5, 2017.

[6] M. Al-Ramahi, O. El-Gayar, and J. Liu, "Discovering Design Principles for Persuasive Systems: A Grounded Theory and Text Mining Approach," in *2016 49th Hawaii International Conference on System Sciences (HICSS)*, 2016, pp. 3074-3083: IEEE.

[7] M. Al-Ramahi, Y. Chang, O. El-Gayar, and J. Liu, "Predicting Big Movers Based on Online Stock Forum Sentiment Analysis," 2015.

[8] C. Noteboom and M. Al-Ramahi, "What are the Gaps in Mobile Patient Portal? Mining Users Feedback Using Topic Modeling," 2018.

[9] A. J. Flanagin and M. J. Metzger, "The role of site features, user attributes, and information verification behaviors on the perceived credibility of web-based information," *New media & society*, vol. 9, no. 2, pp. 319-342, 2007.

[10] M. J. Metzger, "Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 13, pp. 2078-2091, 2007.

[11] A. J. Flanagin and M. J. Metzger, "Digital media and youth: Unparalleled opportunity and unprecedented responsibility," *Digital media, youth, and credibility*, pp. 5-27, 2008.

[12] C. L. Toma, L. C. Jiang, and J. T. Hancock, "Lies in the eye of the beholder: asymmetric beliefs about one's own and others' deceptiveness in mediated and face-to-face communication," *Communication Research*, vol. 45, no. 8, pp. 1167-1192, 2018.

[13] L. Sbaffi and J. Rowley, "Trust and credibility in web-based health information: a review and agenda for future research," *Journal of medical Internet research*, vol. 19, no. 6, 2017.

[14] L. Zhou and Y.-w. Sung, "Cues to deception in online Chinese groups," in *Proceedings of the 41st Annual Hawaii International Conference on System Sciences (HICSS 2008)*, 2008, pp. 146-146: IEEE.

[15] E. J. Briscoe, D. S. Appling, and H. Hayes, "Cues to deception in social media communications," in *2014 47th Hawaii international conference on system sciences*, 2014, pp. 1435-1443: IEEE.

[16] D. S. Appling, E. J. Briscoe, and C. J. Hutto, "Discriminative models for predicting deception strategies," in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 947-952: ACM.

[17] S. Volkova and J. Y. Jang, "Misleading or falsification: Inferring deceptive strategies and types in online news and social media," in *Companion of the The Web Conference 2018 on The Web Conference 2018*, 2018, pp. 575-583: International World Wide Web Conferences Steering Committee.

[18] N. Jindal and B. Liu, "Review spam detection," in *Proceedings of the 16th international conference on World Wide Web*, 2007, pp. 1189-1190: ACM.

[19] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proceedings of the 2008 international conference on web search and data mining*, 2008, pp. 219-230: ACM.

[20] R. Y. Lau, S. Liao, R. C. W. Kwok, K. Xu, Y. Xia, and Y. Li, "Text mining and probabilistic language modeling for online review spam detecting," *ACM Transactions on Management Information Systems*, vol. 2, no. 4, pp. 1-30, 2011.

[21] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, 2011, pp. 309-319: Association for Computational Linguistics.

[22] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*, 2012, pp. 171-175: Association for Computational Linguistics.

[23] V. W. Feng and G. Hirst, "Detecting deceptive opinions with profile compatibility," in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 2013, pp. 338-346.

[24] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, "Detecting product review spammers using rating behaviors," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010, pp. 939-948: ACM.

[25] N. Jindal, B. Liu, and E.-P. Lim, "Finding unusual review patterns using unexpected rules," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010, pp. 1549-1552: ACM.

[26] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 191-200: ACM.

[27] T. Lappas, "Fake reviews: The malicious perspective," in *International Conference on Application of Natural Language to Information Systems*, 2012, pp. 23-34: Springer.

[28] J. Malbon, "Taking fake online consumer reviews seriously," *Journal of Consumer Policy*, vol. 36, no. 2, pp. 139-157, 2013.

[29] H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao, "Spotting fake reviews via collective positive-unlabeled learning," in *2014 IEEE International Conference on Data Mining*, 2014, pp. 899-904: IEEE.

[30] A. Ghose and P. G. Ipeirotis, "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics," *IEEE transactions on knowledge and data engineering*, vol. 23, no. 10, pp. 1498-1512, 2010.