Review Paper

Violence Detection in Social Media-Review

U. Dikwatta and T.G.I. Fernando*

Department of Computer Science, Faculty of Applied Sciences, University of Sri Jayewardenepura, Nugegoda, Sri Lanka

Date Received: 04-11-2019 Date Accepted: 30-11-2019

Abstract

Social media has become a vital part of humans' day to day life. Different users engage with social media differently. With the increased usage of social media, many researchers have investigated different aspects of social media. Many examples in the recent past show, content in the social media can generate violence in the user community. Violence in social media can be categorised into aggregation in comments, cyber-bullying and incidents like protests, murders. Identifying violent content in social media is a challenging task: social media posts contain both the visual and text as well as these posts may contain hidden meaning according to the users' context and other background information. This paper summarizes the different social media violent categories and existing methods to detect the violent content.

Keywords: Machine learning, natural language processing, violence, social media, convolution neural network

1. Introduction

With the rise of Web 2.0, social networking has become a major part of humans' lives in all the aspects including politics, education, health, religion, decision making etc. (Power and Phillips-Wren, 2011). Social media is a way of expressing thoughts freely. It allows people to share information about variety of topics. People share negative or positive thoughts about anything on social media websites. As such social media are computer tools (Siddiqui and Singh, 2016) which allow people to share different media and network with each other. There can be a negative impact on the society because of the way people use social media websites and the poor control of information on these websites. Cyber-bullying, harassment and trolling are some of the offensive actions, perpetrators carrying out using the social media (Pfeffer et al., 2014; Cookingham and Ryan, 2015). Social media is highly used to criticise political matters and these criticisms can be aggressive in some scenarios. Early days, social media was not much popular and only young people gathered around these websites. Nowadays not only young people but also matured and middle-aged people heavily use these websites. A research has done to evaluate the social media impact on USA election in 2016 (Vendemia et al., 2019). This research reported during the time of this election, most of the posts on social media such as Facebook and Twitter are related to politics. With the number of people networking through these websites, number of comments; reviews; posts; images have made social media to surpass the traditional news media such as television and newspapers (Thuseethan and Vasanthapriyan, 2015). Social media websites distribute variety of viewpoints which spread to the community sooner than the other communication methods. As an example, research carried out in (Orrù, 2014) analysed Facebook image posts in Italy on resentment against the immigrants. A specific group created these images; images were shared among individual user accounts within a limited time period. Thuseethan and Vasanthapriyan conducted a research to

*Correspondence: tgi@sjp.ac.lk
Tel: +94 714497227*

analyse the social media as new trend in Sri Lankan digital journalism (Thuseethan and Vasanthapriyan, 2015). They identified six main categories of social media as social networking, bookmarking sites, social news, media sharing, micro blogging and blogs. The impact on social media has risen all over the world and most of the countries are conducting research to find out suitable methods to analyse and classify the content in social media using some mechanisms. One such research (Naher and Minar, 1997) was conducted in Bangladesh to analyse the case studies of the incidents where violence originated from social media websites like Facebook. The recommendation of this research is to use Artificial Intelligence (AI) systems to predict and prevent violent behaviors in social media websites. Popular social media websites are Facebook, Twitter, YouTube, WhatsApp, and Viber. These websites provide number of ways to spread information. One such widely used method is Facebook posts. Image posts are popular among the Facebook and other social media users and most of the images posted are used to criticize the behavior of a person or a group of persons. These posts have been used sarcastically/violently to describe one person's behavior, group or a situation. However in some situations, these posts convey important information to the society as well.

Violence detection in social media can be categorised into areas such as detection of aggregation/hate (Fortuna and Nunes, 2018); cyber-bullying (Salawu et al., 2017); violent images like fire, blood, protest, fight in social media posts (Won et al., 2017). In these categories input is either text or image and output is either binary polarity like aggressive or not; they can have multiple polarities like overtly, covertly and non-aggressive. Some of the recent systems involve with both image and text which have proven to be better results than using one single mode as the input (Lam et al., 2017). The model or the architecture of these systems heavily uses natural language processing and machine learning approaches. Some of the systems use Hidden Markov Model (Rentoumi et al., 2009) and machine learning models like Convolution Neural Networks (CNNs) and Long Short Term Memory (LSTM) networks (Madisetty and Desarkar, 2018). Sentiment analysis and topic classification (Agarwal and Sureka, 2017) approaches have also been used in this task. This paper summarizes and compares these approaches in the context of violence detection in social media.

2. Significance

Several incidents in the recent past have shown the connection between social media content and the real-time violence. Some of the government bodies have discussed this situation with the owners of these websites such as Facebook. In 2016, European commission and four social media companies: Facebook, Twitter, YouTube and Microsoft agreed on a mechanism that, users of these sites can report any hate speech; companies are responsible to assess these requests within 24 hours and remove them immediately if the reported text includes hate. Now Instagram, Google+, Snapchat and other sites also have shown the interest in joining this group.

Many countries have dealt with several social media banns by the year 2019. As an example with the Easter bomb attack, social media sites in Sri Lanka were banned for number of days. Although the sites like Facebook have detection systems, they do not specifically analyse the content which are other than English. In Sri Lanka, native language is Sinhala and most of the posts are created with Sinhala language. The artificial intelligence (AI) systems in these companies which use Natural Language Processing approaches to detect these contents are not capable of detecting Sinhala language. A country like India has more than 10 commonly used languages. Apart from the language, a user can express their feelings in number of different ways like comments, image posts, tags, captions, etc. In order to filter the violent contents all of these need to be considered. To decide on future of automatic violence detection, literature of this area needs to be assessed. Therefore, a survey in this area is highly required at this point of time.

3. Literature Review

Researchers have investigated the content with hate speech, cyber-bullying, violent scenes, fake news and radicalisation in the area of violence detection in social media. These areas can be further divided into violence detection in text, images, videos or combination of these media. These approaches can be divided into machine learning and/or natural language processing based models.

3.1 Violence detection in text data

Violence detection in text data is mainly targets the hate in the text. Hate text is sometimes targets an individual but it impacts on wide group based on gender or race: hate will become a violent incident (Petrocchi et al., 2017). Researchers have used the term "Hate" differently; cyber-bullying, hostile messages and flames are some of the names (Schmidt and Wiegand, 2017). Ellen Spertus developed a system, "Smokey" to detect hostile messages in 1992 as the earliest work in this area to our knowledge (Spertus, 1997). The author used private messages sent through feedback forms of different websites: controversial websites are used as the dataset. Research consists of a natural language processor and a rule-based parser to build a feature vector; decision tree based approach was used to classify the text as 'okay,' 'maybe' or 'flame.' The system was able to classify 98% of 'okay' and 64% of 'flame' correctly. Limitations in the research were difficulties in classifying sarcasm, complex sentences and sentences with mistakes in grammar, punctuation etc. Authors in 2008 (Mahmud et al., 2008) found a similar system which uses a rule-based approach; 'Spertus' is mainly a pattern matching approach while it is based on a general semantic structure that extracts the semantic meaning, the 'Smokey' is message level and system in (Mahmud et al., 2008) is sentence level. As one of the limitations in (Mahmud et al., 2008), researchers found, that the words should be in the lexicon entry to be identified as hate. Dinakar et al. investigated cyber-bullying in social media using textual data (Dinakar et al., 2011). They used an annotation mechanism and measured the agreement between annotators feedback-divided the dataset to binary or multi-class. Data corpus consists of YouTube comments which incorporate with most controversial videos. Binary polarity gives better results than multi-class polarity. They have used Naïve Bayes, Rule-Based, Decision Trees and SVM classifiers; Rule-Based mechanism gives more accurate results-SVM classifier gives more reliable results with regard to the data with higher agreement level. Basave et al. have introduced a violence detection model (VDM) in 2013 to detect violent documents in social media based on the word prior knowledge that detects what words are violent and non-violent (Basave et al., 2013). This model does not require labeled data to train the model. The proposed method introduced an entropy based word prior knowledge gaining approach. The proposed method is compared with baseline methods like Joint-Sentiment Topic (JST) and the Partial Labeled LDA (PLDAL): the proposed method gives better results in detecting violent and non-violent content. Petrocchi and Tesconi proposed a text classification method in Italian Facebook comments (Petrocchi et al., 2017). Hate levels are 'No Hate', 'Weak Hate' and 'Strong Hate'; they categorized hate comments to religion, physical and/or mental handicap, socio-economic status, politics, race, sex and gender issues, and other. They have measured the agreement between annotations on a given task. Depending on the agreement, they have divided the hate level to two classes (merged strong and weak hate) or three classes which is similar to the approach in (Dinakar et al., 2011). Long Short Term Memory (LSTM) and Support Vector Machine (SVM) are used as the approaches. Both classifications give better results with the dataset that considered the annotator agreement.

3.2 Violence detection in images

Wang et al. have detected violent images using their own dataset which they composed by querying keywords through search engines (D. Wang et al., 2012). Group of annotators annotated images as violent, neutral and non-violent. Experiment focused on violent and non-violent categories.

They have used Bag of Words (BoW) model with five different features (Scale Invariant Transform (SIFT), Histogram of Oriented Gradient (HOG), colour histograms, and Local Binary Patterns (LBP). Results show that the features for text analysis have better results than other four features. Won et al. have studied protest activities and measured the violence in social media images (Won et al., 2017). They have created their own social media dataset with protest activities and peaceful images mainly from Twitter. Amazon Mechanical Turk annotators annotated the images depending on the visual appearance. They fed part of the dataset to Residual Network (ResNet) with 50 layers and tested the model with the other part of the dataset. They compared the results with the manual annotation: they calculated Pearson's correlation coefficients and R^2 values. They observed a high correlation between the manual and results generated by the model. Apart from measuring the violence they trained the model: categorize images as protest and non-protest, identify different visual attributes, analyze sentiment and used another CNN to capture the emotions of humans in images. Sun et al. have detected violent behaviours of still images using multi-view maximum entropy discriminant (MVMED+) with different views (Sun et al., 2019). This approach involves multiple feature sets and each feature set is referred as a view. They developed "violence image recognition" (VIR) database: according to the researchers VIR is the largest open dataset with violent images and they have collected the images by querying search engines. Features can be categorized as high-level, mid-level and low-level features. Low-level features are the pixel intensities or colours. Mid-level features are the object detection, image annotations and other similar mechanisms. High-level features are used to image classification and clustering tasks. In this research, low-level features are considered in order to identify violent images. The paper concludes that the proposed method gives better results than CNN, Naive Bayes (NB) and k-nearest neighbour (KNN).

3.3 Violence detection in videos

As the first step system extracts the key elements like frames and key frames. As the second step system extracts features like bag of local features and colour histograms. Valle et al. have used a Support Vector Machine (SVM) as the classifier (Valle et al., 2012). The classifier gives votes to each feature type and depending on the majority of votes, a video gets a label. They have done three tests with three datasets to detect pornography, violence and legitimate videos. The conclusion of this research is that spatio-temporal bag of features give better results than other features for all three datasets. This has shown higher results in violence detection. Souza et al. have also found that spatio-temporal features produce good results in identifying violent videos (De Souza et al., 2010).

Sudhakaran and Lanz have studied violent and non-violent videos generated from surveillance cameras (Sudhakaran and Lanz, 2017). CNN alone can identify spatial information of a given video frame; video consists of both spatial and temporal data: RNN is a better network to identify temporal information. They have used Convolutional Long Short Term Memory (ConvLSTM) architecture that can identify both spatial and temporal features of a video. The proposed method is compared against several baseline methods against hockey fight, movies (Bermejo Nievas et al., 2011) and violent-flows crowd violence (Hassner et al., 2012) dataset and it has given better results against hockey and movies datasets. They obtain second best results for violent-flow dataset. They have compared the results with LSTM model and the accuracy is better in the proposed network and less number of parameters involve in the proposed method than the LSTM method.

3.4 Multimodal violence detection

Most of the violent content includes several modalities like audio, visual and text. Therefore, researchers investigate methods to use multiple features that include audio, visual and text to detect violent content: these approaches are named as multimodal violence detection.

Morency and Mihalcea proposed the first model to decide sentiment in videos using text, audio and visual features (Morency et al., 2011). They used Hidden Markov based models to classify the images and compared the model with text only, visual only and audio only approaches: obtained highest results for the proposed method in F1 score, Precision and Recall. You et al. analysed sentiment using both visual and textual elements of tweets (You et al., 2015). They have used a Convolution Neural Network (CNN) model to analyse visual features and unsupervised language model to analyse text features. They have tested the model on Getty and Twitter images. Results are compared only with textual data and visual data and calculated the precision, Recall, F1 score and Accuracy. Textual data only gives the highest precision and proposed method gives higher Recall, F1 and Accuracy. Visual data only gives comparable good results. Baecchi et al. have proposed an unsupervised learning method using Continuous Bag of Words (CBOW) and a neural network model Denoising Autoencoder (Baecchi et al., 2016). Denoising Autoencoder is a slightly different version of Autoencoder, which enables to input an image different than the original image and it generalizes the model without tightly fitting into the input dataset. A sliding window mechanism is applied over the tweet and CBOW captures the text and Autoencoder captures the image part: a local polarity score is calculated to each window and summed up to get a final score. The proposed method is compared against two more methods and five datasets: the proposed method has given the best results. Proposed method is compared against text, images and combination of texts and images against previous two baseline methods, and the proposed method has given the best results.

4. Methodologies Involved with Violence Detection

Machine learning approaches

Supervised, semi-supervised and unsupervised

In (Xiang et al., 2012) Xiang et al. used an unsupervised approach to detect tweets related to offensive topics. The model use topic, lexicon features and supervised classification algorithms. Several supervised learning methods were used; Decision Tree Learning, Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest (RF). LR performs slightly better results than other supervised algorithms and LR has given 5.4% improvement over keyword matching baseline method. Research in (Putri, 2019) was done in the context of Indonesia news sites to identify hoax content. After several pre-processing steps, several machine learning algorithms have been applied to compare the efficiency of results. Algorithms used here are Multilayer Perceptron, Support Vector Machine, Naïve Bayes, Random Forest, and Decision Tree algorithms. The results are compared with F1-score, Recall and Precision. According to the results obtained, Random Forest has given the best results.

When identifying radical opinion in web, supervised learning which require pre-labelling is difficult: semi-supervised learning can be a good approach to identify radical opinion. Yang and Chen (Yang and Chen, 2012) proposed a system based on unlabeled data using Support Vector Machine (SVM) and Naïve Bayes (NB): results shows labelling unlabeled data gives better results. The semi-supervised cross feature learning algorithm (SCFL) is used in (Gong et al., 2008) to detect violent scenes in movies using auditory and visual features. The results are compared with supervised learning mechanism support vector machine (SVM) and obtained better results.

Magistry et al. (2016) have proposed an unsupervised system to detect sentiment in microblogs and the results showed average F1 score of 87.2%. The research is carried out on Chinese language. Abdelfatah et al. (2017) have proposed a method to detect violent content in Arabic social media using unsupervised learning. They have used unsupervised dimensionality reduction (DR) method and unsupervised cluster algorithm. They have used k-means as the algorithm and two DR techniques used are Principal Component Analysis (PCA) and Sparse Gaussian Process Latent Variable Model (SGPLVM). Results show that SGPLVM with K-means works better than PCA. Researchers have compared classical approaches like SVM with deep learning approaches using nine available data sources (Chen et al., 2018). They have found that when the classes are imbalanced, deep learning approaches give better results and when oversampling is used to balance the classes, SVM gives better results.

Deep learning

A majority voting based ensemble method is used with convolution neural network (CNN), long short-term memory (LSTM) and Bi-directional long short-term memory (Bi-LSTM) to identify overtly, covertly and non-aggressive text in Facebook posts. Results obtained are higher in non-aggressive class than other two classes (Madisetty and Desarkar, 2018). Aroyehun and Gelbukh have proposed a similar method that compared Native Bayes Support Vector Machine (NBSVM) and other deep neural networks (Aroyehun and Gelbukh, 2018). Deep learning methods to outperform NBSVM, data augmentation, pseudo labelling and sentiment scores are used. They obtained 5% weighted macro improvement with the data augmentation and 7% weighted macro improvement with data augmentation and pseudo labelling over NBSVM.

Natural language processing approaches

Schmidt and Wiegand (2017) have summarised the available research on natural language processing for hate speech detection. Researches have used natural language processing methods as feature selection method in text and supervised learning is applied as the classification method (Davidson et al., 2017; Malmasi and Zampieri, 2017). In (Malmasi and Zampieri, 2017), they have used natural processing feature selection algorithms like character n-gram, word n-gram, skip-gram and brown clusters. As supervised learning classifier Support Vector Machine (SVM) is used. Results shows character 4-gram as the best feature with 78% accuracy. In (Davidson et al., 2017), researchers used bigram, unigram and trigram with TF-IDF method. They tested the dataset with different supervised learning algorithms and found Logistic Regression works better. Best model has given overall precision of 0.91, recall of 0.90 and F1 score of 0.90.

Other approaches

Some of the earliest approaches involve with hate speech detection are rule-based (Spertus, 1997, Mahmud et al., 2008). These approaches failed to detect complex sentences and depend on the keywords in the system. Researchers have used keyword search as a web mining technique to filter hate groups (Ting et al., 2013).

5. Text and Object Detection in Images

In the analysis of social media image posts, identifying and detection of texts and objects in the image posts is critical. Text detection and text recognition are the main steps followed in (Wang et al., 2012). Convolution neural network with non-maximal suppression is used to get the final result. Bounding box mechanism was used to detect the text. Performance of text recognizer is higher than the text detector. Research in (Zou et al., 2019) is a survey paper that consists of classical and recent approaches to detect objects in images. With the usage of convolution neural networks (CNN), one stage object detection algorithms like You Only Look Once (YOLO), Single Shot Multibox Detector (SSD) and RetinaNet are used as object detection methods.

6. Text Embedding

In (Gomez et al., 2019; Modha and Majumder, 2019), a mechanism has been proposed to embed text as the input of deep learning networks. As the two research states, available embeddings are Latent Dirichlet Allocation (LDA), Word2Vec, FastText, Doc2Vec and GloVe. In (Modha and Majumder, 2019), researchers evaluate bag of words, TF-IDF, Word2Vec, FastText, Glove and Doc2Vec in classifying aggressive text in social media. The conclusion of this paper related to embedding is that FastText performs better than Glove, Word2Vec and Doc2Vec: FastText is based on character n-gram.

7. Violence Detection Sinhala Language

A research was conducted to analyse racist and non-racist comments in Sri Lankan context. A two-class support vector machine mechanism was used to train the network with two sets as racist and non-racist comments obtained from Facebook (Dias et al., 2018). Dataset was pre-processed using n-gram features and two-class support vector machine was applied to the 75% of training data in the dataset. But the precision of the results in this model dropped at some point. Another research was conducted to compare different mechanisms to identify hate speech in a local English dataset (Ruwandika and Weerasinghe, 2018). The dataset consists of English comments published in a news site of Sri Lanka. Different classifiers with different features were evaluated and compared. Five machine learning models which were used to achieve the task are Support vector machines, Logistic Regression algorithm, Naïve Bayes algorithm, Decision Tree algorithm and K-Means clustering algorithm. Bag of words, TF-IDF and two more feature types are used as features. Google bad word list ("Full List of Bad Words and Swear Words Banned by Google," 2019) was used as the hate lexicon to extract the features from the dataset. As a result of the research, Naive Bayes and TD-IDF output a high F-score value and supervised learning models worked better than the unsupervised learning models.

8. Concluding Remarks

According to the literature, results differ depending on the datasets. But according to some of the research, deep learning methods have outperformed classical machine learning methods like SVM. Character n-gram method gives better performance when classifying text. As the literature suggests, most of the research are based on social media text or images. Although researchers have done research on multimodal content: caption of the images and image itself have been taken into consideration. Recently social media image posts which contain both text and objects are a popular way of communicating. We need to address this problem. Deep learning approaches with word embedding and object detection algorithms like YOLO can give better performance in this task. First task is the creation of balance (negative and positive) data corpus. Then text and object detection along with text embedding should be followed. Finally, classification algorithm should be decided.

References

Agarwal, S., Sureka, A., 2017. Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on Tumblr Micro-Blogging website.

Aroyehun, S.T., Gelbukh, A., 2018. Aggression detection in social media: using deep neural networks, data augmentation, and pseudo labeling, 1st Workshop on trolling, aggression and cyberbullying (TRAC-1). Santa Fe, USA, 8.

Baecchi, C., Uricchio, T., Bertini, M., Del Bimbo, A., 2016. A multimodal feature learning approach for sentiment analysis of social network multimedia. *Multimedia Tools and Applications*, 75:2507-2525.

Basave, A.E.C., He, Y., Liu, K., Zhao, J., 2013. A weakly supervised bayesian model for violence detection in social media: IJCNLP, 109-117.

Bermejo Nievas, E., Deniz Suarez, O., Bueno García, G., Sukthankar, R., 2011. Violence detection in video using computer vision techniques. *Computer Analysis of Images and Patterns*, 332-339.

Chen, H., McKeever, S., Delany, S.J., 2018. A comparison of classical versus deep learning techniques for abusive content detection on social media sites. *Social Informatics*, 117-133.

Cookingham, L.M., Ryan, G.L., 2015. The impact of social media on the sexual and social wellness of adolescents. J*ournal of Pediatric and Adolescent Gynecology*, 28:2-5.

Davidson, T., Warmsley, D., Macy, M., Weber, I., 2017. Automated hate speech detection and the problem of offensive language, 11th International AAAI conference on web and social media: association for the advancement of artificial intelligence (AAAI), 4.

De Souza, F.D.M., Cha, G.C., Do Valle, E.A., De A Araujo, A., 2010. Violence detection in video using spatio-temporal features, 23rd SIBGRAPI conference on graphics, patterns and images: IEEE, Gramado, 224-230.

Dias, D.S., Welikala, M.D., Dias, N.G.J., 2018. Identifying racist social media comments in sinhala language using text analytics models with machine learning, 18th International conference on advances in ICT for emerging regions (ICTer): IEEE, Colombo, Sri Lanka, 1-6.

Dinakar, K., Reichart, R., Lieberman, H., 2011. Modeling the detection of textual cyberbullying, 5th International AAAI conference on weblogs and social media (SWM'11): Barcelona, Spain, 7.

E Abdelfatah, K., Terejanu, G., A Alhelbawy, A., 2017. Unsupervised detection of violent content in arabic social media, computer science and information technology (CS and IT). 4th International conference on computer science and information technology: Academy and Industry Research Collaboration Center (AIRCC), 1-7.

Fortuna, P., Nunes, S., 2018. A survey on automatic detection of hate speech in text. ACM Computing Surveys, 51:1-30.

Full List of Bad Words and Swear Words Banned by Google, 2019. Free Web Headers. Available at: https://www.freewebheaders.com/full-list-of-bad-words-banned-by-google/ (Accessed April 2019).

Gomez, R., Gomez, L., Gibert, J., Karatzas, D., 2019. Learning to learn from web data through deep semantic embeddings. *Computer Vision-ECCV 2018 Workshops*, 514-529.

Gong, Y., Wang, W., Jiang, S., Huang, Q., Gao, W., 2008. Detecting violent scenes in movies by auditory and visual cues. *Advances in Multimedia Information Processing-PCM 2008,* 317-326.

Hassner, T., Itcher, Y., Kliper-Gross, O., 2012. Violent flows: real-time detection of violent crowd behavior, IEEE computer society conference on computer vision and pattern recognition workshops: IEEE, Providence, RI, USA, 1-6.

Lam, V., Phan, S., Le, D.D., Duong, D.A., Satoh, S., 2017. Evaluation of multiple features for violent scenes detection. *Multimedia Tools and Applications*, 76:7041–7065.

Madisetty, S., Desarkar, M.S., 2018. Aggression detection in social media using deep neural networks, 1st Workshop on trolling, aggression and cyberbullying (TRAC-1): Santa Fe, USA, 8.

Magistry, P., Hsieh, S.-K., Chang, Y.Y., 2016. Sentiment detection in micro-blogs using unsupervised chunk extraction. *Lingua Sinica*, 2.

Mahmud, A., Zubair, K., Mumit, K., Ahmed, 2008. Detecting flames and insults in text, 6th International conference on natural language processing (ICON' 08): 10.

Malmasi, S., Zampieri, M., 2017. Detecting Hate Speech in Social Media.

Modha, S., Majumder, P., 2019. An empirical evaluation of text representation schemes on multilingual social web to filter the textual aggression.

Morency, L.P., Mihalcea, R., Doshi, P., 2011. Towards multimodal sentiment analysis: harvesting opinions from the web, 13th International conference multimodal interfaces: ACM, 169-176.

Naher, J., Minar, M.R., 1997. impact of social media posts in real life violence: a case study in Bangladesh, 8.

Orrù, P., 2014. Racist discourse on social networks: a discourse analysis of facebook posts in Italy, 21.

Petrocchi, M., Tesconi, M., Dell'Orletta, F., Cimino, A., Del Vigna, F., 2017. Hate me, hate me not:Hate speech detection on Facebook, 1st Italian conference on cybersecurity: 86-95.

Pfeffer, J., Zorbach, T., Carley, K.M., 2014. Understanding online firestorms: Negative word-of-mouth dynamics in social media networks. *Journal of Marketing Communications*, 20:117-128.

Power, D.J., Phillips-Wren, G., 2011. Impact of social media and web 2.0 on decision-making. *Journal of Decision Systems* 20, 249-261.

Putri, T.T.A., 2019. Analysis and detection of HOAX contents in Indonesian news based on machine learning, 4:8.

Rentoumi, V., Giannakopoulos, G., Karkaletsis, V., Vouros, G.A., 2009. Sentiment analysis of figurative language using a word sense disambiguation approach, International conference RANLP. association for computational linguistics: Borovets, Bulgaria, 370-375.

Ruwandika, N.D.T., Weerasinghe, A.R., 2018. Identification of hate speech in social media, 18th international conference on advances in ICT for emerging regions (ICTer). 2018, IEEE, Colombo, Sri Lanka, 273-278.

Salawu, S., He, Y., Lumsden, J., 2017. Approaches to automated detection of cyberbullying: A Survey. IEEE Transactions on Affective Computing 1-1.

Schmidt, A., Wiegand, M., 2017. A survey on hate speech detection using natural language processing, 5th International workshop on natural language processing for social media: Association for Computational Linguistics, Valencia, Spain, 1-10.

Siddiqui, S., Singh, T., 2016. Social media its impact with positive and negative aspects. *International Journal of Computer Applications Technology and Research*, 5:71-75.

Spertus, E., 1997. Smokey automatic recognition of hostile messages, Innovative Applications of Artificial Intelligence (IAAI): American Association for Artificial Intelligence, 1058-1065.

Sudhakaran, S., Lanz, O., 2017. Learning to detect violent videos using convolutional long short-term memory.

Sun, S., Liu, Y., Mao, L., 2019. Multi-view learning for visual violence recognition with maximum entropy discrimination and deep features. *Information Fusion*, 50:43-53.

Thuseethan, S., Vasanthapriyan, S., 2015. Social media as a new trend in sri lankan digital journalism: a surveillance. *Asian Social Science*, 11.

Ting, I.H., Wang, S.L., Chi, H.M., Wu, J.S., 2013. Content matters: a study of hate groups detection based on social networks analysis and web mining, IEEE/ACM international conference on advances in social networks analysis and mining-ASONAM'13: ACM Press, Niagara, Ontario, Canada, 1196-1201.

Valle, E., de Avila, S., de Souza, F., Coelho, M., Araújo, A., 2012. Content-Based Filtering for Video Sharing Social Networks, Brazilian symposium on information and computer system security (SBSeg): 625-638.

Vendemia, M.A., Bond, R.M., DeAndrea, D.C., 2019. The strategic presentation of user comments affects how political messages are evaluated on social media sites: Evidence for robust effects across party lines. *Computers in Human Behavior*, 91:279-289.

Wang, D., Zhang, Z., Wang, W., Wang, L., Tan, T., 2012. Baseline results for violence detection in still images, IEEE 9th international conference on advanced video and signal-based surveillance (AVSS): IEEE, Beijing, China, 54-57.

Wang, T., Wu, D.J., Coates, A., Ng, A.Y., 2012. End-to-end text recognition with convolutional neural networks, in: pattern recognition (ICPR), 21st international conference: IEEE, 3304-3308.

Won, D., Steinert-Threlkeld, Z.C., Joo, J., 2017. Protest activity detection and perceived violence estimation from social media images.

Xiang, G., Fan, B., Wang, L., Hong, J., Rose, C., 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus, 21st ACM international conference on information and knowledge management-CIKM '12: ACM Press, Maui, Hawaii, USA, 1980.

Yang, M., Chen, H., 2012. Partially supervised learning for radical opinion identification in hate group web forums, IEEE international conference on intelligence and security informatics (ISI 2012): IEEE, Washington, DC, USA, 96-101.

You, Q., Luo, J., Jin, H., Yang, J., 2015. Joint visual-textual sentiment analysis with deep neural networks, 23rd ACM international conference on multimedia-MM'15: ACM Press, Brisbane, Australia, 1071-1074.

Zou, Z., Shi, Z., Guo, Y., Ye, J., 2019. Object detection in 20 years: a survey.