D-FJ: Deep Neural Network Based Factuality Judgment

Ankan Mullick
ankan.mullick@infoedge.com
Infoedge
India

Sourav Pal
souravpal.2014@iitkgp.ac.in
IIT Kharagpur
India

Projjal Chanda
projjalchanda@iitkgp.ac.in
IIT Kharagpur
India

Arijit Panigrahy
arijitpanigrahy@iitkgp.ac.in
IIT Kharagpur
India

Anurag Bharadwaj
chatrasen@iitkgp.ac.in
IIT Kharagpur
India

Siddhant Singh
siddhantsingh@iitkgp.ac.in
IIT Kharagpur
India

Tanmoy Dam
t.dam@student.adfa.edu.au
The University of New South Wales
Australia

ABSTRACT

In the modern digital world, people are becoming polarized day by day and expressing their emotions in online digital world. The posted sentences in the online platform, can be categorized as factual or opinionated. Factuality of a sentence can be proved as true or false but opinion is an ornament on top of a fact. Factuality is also the measurement metrics to identify the credibility of an online text. So identifying factuality and someone's opinion of the posted text is an important problem. But, people use to post so much stuff online as automatic factuality detection with higher accuracy of an article is very relevant nowadays.

In this paper we have developed a deep neural network based factuality judgment model (D-FJ). The first step of our approach is to develop a two class classifier model to detect factual and opinionated sentences from online news media. We have also shown how factuality, opinionatedness and sentiment fraction of different news articles changes over certain events in different time frames. Extensive evaluations show that our proposed model provides better results than the existing models in terms of precision, recall and accuracy than existing approaches.

ACM Reference Format:
Ankan Mullick, Sourav Pal, Projjal Chanda, Arijit Panigrahy, Anurag Bharadwaj, Siddhant Singh, and Tanmoy Dam. 2019. D-FJ: Deep Neural Network Based Factuality Judgment. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Online virtual world is a large domain for user engagement and billion dollar business through advertisement. People are engaging themselves in online world through commenting, sharing, liking pages etc. Depending on the nature of the articles for different sections, user engagement varies - for example: users mostly want more facts in case of science, technology sections but highly opinionated articles (less factual) in case of politics, editorial sections. Thus, factuality dynamics is an important factor to determine the future user engagements and online revenues. News agencies are interested to know the demography of the factuality dynamics of their articles along with sentiments variations and how the dynamics vary depends on certain events. Factuality detection and factuality dynamics change are relevant problems to solve for tuning the revenue parameters in digital world.

For the last two decades, people have engaged themselves working on factuality and opinion mining. FactBank was created by Sauri and Pustejovsky [18]. Marneffe et al. [5] evaluated FactBank ratings. A two-dimensional polarity and certainty based factuality annotation scheme was built by Sauri and Pustejovsky [19]. Soni et al. [20] detect factuality of tweet text using predicates (cues) and groups [17]. Asher et al.[2] categorizes opinions. Qadir [14] detects opinion and fact sentences specific to product features in customer reviews. Kim and Hovy [7] extract opinions, opinion holders from news text. Yu and Hatzivassiloglou [21] separate facts from opinions. Liu [8] describes opinion mining and sentiment analysis. Agarwal et al. [1] do sentiment analysis of Twitter data. Fang and Zhan [6] analyze product review sentiments. Baly et al. [3] focus on predicting the factuality of reporting and bias of news media. Rudinger et al [16] build neural network model for event factuality prediction. Rajkumar et al. [15] develop opinion fact identifier algorithm in a graphical framework. Mullick et al. [11] extend the graphical framework adding diversity algorithm to compute diverse opinion and fact. Mullick et al. [10, 12] build a generic opinion-fact classifier for classical and social media texts[1]. This also shows opinionatedness and factuality distribution across different news article sections.

But none of these work focused on factuality dynamics and how it changes over time and specific event along with the changes of sentiment patterns in online news. We have built a deep neural

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] We have used [15] and [12] methods as baselines.

Table 1: Subsection and Events

Subsection	Events
Technology	Iphone X, Google Keynote, Amazon Echo, Nasa Mars Mission
Sports	Football World Cup 2014, Olympics 2016, British GP 2017, Neymar's transfer to PSG
Politics	Indian General Election 2014, Brexit, US Presidential Election 2012, France Presidential Election
World	Rohingya Crisis, Las Vegas Shooting, Catalonia Referendum, Hurrican Irma
Finance	Greece Debt crisis, India's Demonetization, Zimbabwe's dwindling currency, UK Inflation

Table 2: Test Dataset statistics

Used	Number of Docs	Avg. Length of an article
Dataset	Event/Month	Event/Month
Technology	50/173	37.4/39.4
Sports	50/194	34.3/32.2
Politics	50/186	35/35.6
World	50/189	40.1/39.4
Finance	50/190	36.2/33.8

network approach to identify factuality better in terms of accuracy, precision, recall. Our model detects change in sentiment and opinionatedness-factuality w.r.t. time and event.

2 DATASET

We have used two public datasets Mullick et al. [12] - a) standard Multi-Perspective Question Answering (MPQA) (contains 535 documents) and b) 120 news articles crawled from Yahoo news. Each document is a news article pertaining to some topic. Each sentence is already labelled as opinion ("O") or factual ("F").

For further experimentation, we have crawled news articles from 'The Guardian'. The articles were collected from a set of 5 subsections namely (a) Politics, (b) World, (c) Sports, (d) Technology and (e) Finance. We have crawled articles based on two levels (a) Event specific dataset (b) One Month dataset. For event specific article section (like 'Politics', 'Sports' etc.), we choose 5 prominent events from last 10 years such as FIFA World cup 2010 for 'Sports' or the great US recession for 'Finance'. The details for the events each subsection is present in Table 1. For each of the chosen events, we crawled 10 articles from the Guardian website. The sentences of each of the crawled articles were annotated manually for calculating Precision and Recall. For month specific dataset, we crawled the articles of the aforementioned subsections of a specific month. In our case, we built the dataset from the month of September 2017. The statistics are present in Table 2.

3 EXPERIMENT

From the 535 documents in the MPQA dataset and 120 Yahoo datasets, we computed several features (1) POS Tag based features[2] - number of nouns, adverbs etc.; (2) Dependency based features[3]: presence of adjective modifier etc.[4] and (3) Others: (a) count of the strong polar words, weak polar words in the sentence (b) polarity of the root verb of the sentence, (c) opinionated n-grams (d) presence of modal verbs, (e) opinionated and factual words. Based on these features, we have developed various classification models to detect fact sentences, such as Naive Bayes classifier (NB-R) [15], Decision Tree (DT), JRip classifier, Multilayer perceptron (MLP), Support Vector Machine (SVM), Random Forest (RF), Bagging with Random Forest [12] (BG+RF) classifiers on the MPQA and Yahoo datasets. The results are shown in Table 3.

Table 3: Precision, Recall, Accuracy based on different standard classification models

	MPQA	MPQA	MPQA	Yahoo	Yahoo	Yahoo
Model	P	R	A	P	R	A
NB-R	67.12	57.54	58.40	60.87	61.74	60.61
DT	65.91	66.23	70.003	57.75	61.43	60.63
MLP	68.21	63.23	69.10	64.72	65.10	66.56
SVM	53.23	52.98	53.12	64.62	50.47	63.97
JRip	74.10	71.05	70.11	63.06	70.02	70.17
RF	58.21	59.34	59.13	69.31	70.45	70.95
BG+RF	74.41	75.51	74.73	76.16	74.34	73.19

On these two datasets, (a) Recurrent Neural Network (RNN) and (b) Long Short Term Memory (LSTM) models were built using Glove embedding word vectors[5] Batch size parameter was tuned for better Accuracy, Precision and Recalls. Batch size was varied from 32 to 100 and results were computed for different values in the range. The aim was to select the best model based on these criteria for the next phase. We have used different activation functions - 'Sigmoid', 'Tanh' and 'Relu' for LSTM and RNN are shown in Table 4. We observed that deep neural network based models have better performance than classical models. From the table 3 and 4, best results are obtained with batch size = 32 and for 'Sigmoid activation functions' for LSTM in terms of precision, recall and accuracy. [6]. Thus, we have used this LSTM model to detect factuality from articles.

We have also crawled one year articles (2017 January - December) from 'The Guardian' news articles for five sections - World, Technology, Sports, Politics and Finance. Using the above LSTM model, we have calculated factuality of the articles year-wise, for a particular month (September, 2017) and event-wise[7].

From the Table 5, it is shown that - technology and finance sections are highly factual but sports, politics are less factual and world section's factuality is nearly half of its content. Table 5 infers that though for a particular month factuality decreases across all sections but in case of even-wise factuality behaviour changes -

[2]Standford POS tagger [9] has been used to identify the tags.

[3]Stanford Dependency Parser [4] has been used to identify the dependencies.

[4]Details of the features are in [12].

[5]Pre-trained Glove embedding word vectors were used. GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

[6]We have fine tuned the dropout parameter for better results.

[7]factuality score is the average of individual scores for articles of a particular section.

Table 4: DNN with Glove Vector Models on MPQA

			MPQA	MPQA	MPQA	Yahoo	Yahoo	Yahoo
Model	Activation	Batch Size	P	R	A	P	R	A
LSTM	Tanh	32	78.10	86.05	79.23	69.32	73.27	67.91
LSTM	Sigmoid	32	80.03	89.31	80.19	77.31	79.28	74.08
LSTM	Relu	32	79.16	82.98	78.82	66.54	69.38	65.72
RNN	Tanh	32	69.03	74.50	71.24	67.14	66.43	64.09
RNN	Sigmoid	32	77.21	84.93	76.30	78.87	71.10	64.58
RNN	Relu	32	75.78	76.17	73.28	68.81	64.34	69.85

for high and medium factual sections, (technology, finance, world) factuality decreases (w.r.t year wise distribution) but for less factual sections (sports, politics), factuality increases. Event-wise factuality increases across all sections w.r.t month's distributions. Event-wise factuality changes in big margin for less factual sections mostly (w.r.t month and year)- like sports, politics.

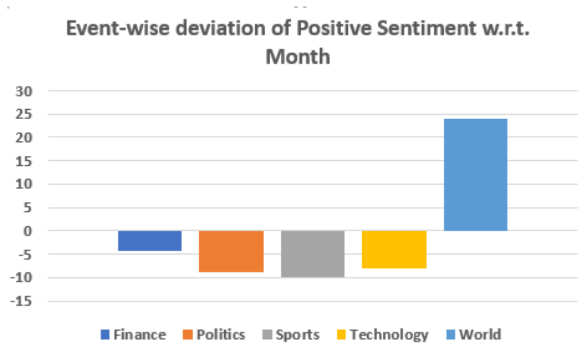

Figure 1: Event-wise positive sentiment deviation of factuality w.r.t. month

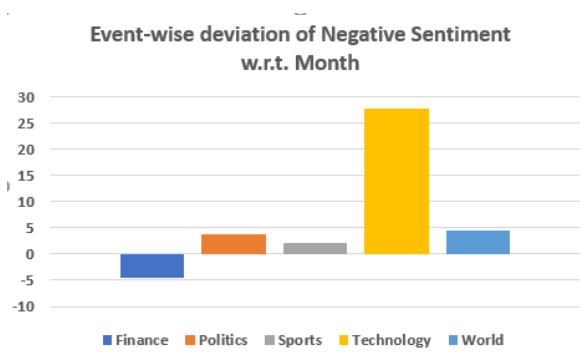

Figure 2: Event-wise negative sentiment deviation of factuality w.r.t. month

We have experimented how article-sentiment varies for events with respect to the month. Figure 1 refers that positive article-sentiment level decreases for any events across finance, politics, sports and technology sections but increases only for world sections. Deviation for finance is less significant and for world is most

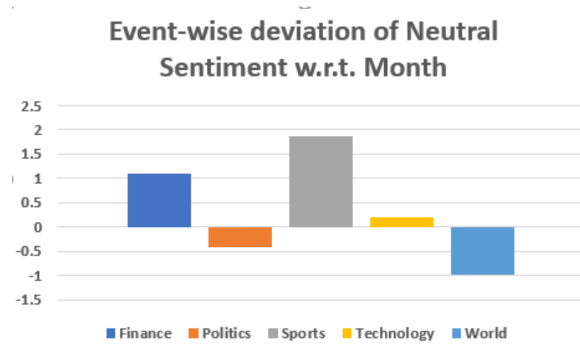

Figure 3: Event-wise neutral sentiment deviation of factuality w.r.t. month

significant compared to others. Figure 2 infers that for any new events, negative article-sentiments also increases for politics, sports, technology and world section but decreases for finance section. Negative sentiment deviation for technology is the most - this is due to the fact that people are often talk about the side effects or harmfulness when a new technology articles is emerged by the online publishing house. Figure 3 shows the event specific deviations for neutral comments across all sections - finance, sports and technology sections are positively deviated but politics and world are negatively deviated. But percentage of deviation is non-significant (~ 0-1%) for most of the sections. Only sport sections has positive deviation of 1.8%. This is due to the fact that people often take a particular stance (polarity) on a certain event so neutral factuality deviation is less.

4 CONCLUSION

We have proposed a deep neural network based factuality detection algorithm to identify the factual dynamics of online news sections. We have also identify the event-wise factuality, positive and negative sentiment deviation across different news article sections. There are some aspects of the dynamics need to be experimented more like - (a) how different categories of factuality [20] varies with news article sections. (b) factuality dynamics of the comment sections. (c) one complex or compound sentence can contain multiple facts or opinions or one sentence may contain multiple factuality or opinionated phrases. To understand the dominant phrase or sentence that determine the overall sentence characteristics. There are several future scopes of this work - (1) currently there is no

Table 5: Opinion Fraction of Guardian Dataset

Subsection	Yearly	Monthly, Deviation (%)	Event-Wise, Yearly Deviation (%), Monthly Deviation (%)
World	0.49	0.45, +8.16	0.48, +4.1, -6.67
Tech	0.73	0.52, +28.77	0.69, +5.48, -32.7
Sports	0.47	0.39, +17.02	0.49, -4.26, -25.64
Politics	0.30	0.32, +6.67	0.38, -26.67, -18.75
Finance	0.67	0.53, +20.9	0.62, +7.46, -16.98

mathematical model that defines factuality dynamics and its deviations with characteristics. Our immediate future step is to build the factuality model (2) increase size of the annotated datasets and make a bigger one, than the current one then deep neural network would have worked better. (3) It will also be quite informative to examine how sentiment and factuality vary demographically (sex, age, region etc.), for different time frames like days of week (weekdays vs weekends), monthly (start of the month vs end of the month) or hourly (morning vs work hours vs evening vs late night). (4) We can explore different characteristics of LIWC (Linguistic Inquiry and Word Count) [13] and how it is related to the factuality. (5) developing a chat moderator to automatically recommend or highlight important facts about the article to the users for encouraging commenting against top ranked facts or re-commenting against a comment related to a fact.

REFERENCES

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*. 30–38.

[2] Nicholas Asher, Farah Benamara, and Yvette Yannick Mathieu. 2009. Appraisal of opinion expressions in discourse. *Lingvisticæ Investigationes* 32, 2 (2009).

[3] Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. *arXiv preprint arXiv:1810.01765* (2018).

[4] Marie-Catherine De Marneffe and Christopher D Manning. 2008. *Stanford typed dependencies manual*. Technical Report. Stanford University.

[5] Marie-Catherine De Marneffe, Christopher D Manning, and Christopher Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational linguistics* 38, 2 (2012), 301–333.

[6] Xing Fang and Justin Zhan. 2015. Sentiment analysis using product review data. *Journal of Big Data* 2, 1 (2015), 5.

[7] Soo-Min Kim and Eduard Hovy. 2006. Extracting Opinions, Opinion Holders, and Topics expressed in Online News Media Text. In *Workshop on Sentiment and Subjectivity in Text*. 1–8.

[8] Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies* 5, 1 (2012), 1–167.

[9] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *ACL: Demo*.

[10] Ankan Mullick, Surjodoy Ghosh D, Shivam Maheswari, Srotaswini Sahoo, Suman Kalyan Maity, Pawan Goyal, et al. 2018. Identifying Opinion and Fact Subcategories from the Social Web. In *ACM GROUP*.

[11] Ankan Mullick, Pawan Goyal, and Niloy Ganguly. 2016. A graphical framework to detect and categorize diverse opinions from online news. In *PEOPLES*.

[12] Ankan Mullick, Shivam Maheshwari, Soumya C, Pawan Goyal, and Niloy Ganguly. 2017. A Generic Opinion-Fact Classifier with Application in Understanding Opinionatedness in Various News Section. In *WWW Companion*.

[13] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71 (2001), 2001.

[14] Ashequl Qadir. 2009. Detecting opinion sentences specific to product features in customer reviews using typed dependency relations. In *Workshop on Events in Emerging Text Types*. ACL.

[15] Pujari Rajkumar, Swara Desai, Niloy Ganguly, and Pawan Goyal. 2014. A Novel Two-stage Framework for Extracting Opinionated Sentences from News Articles. *TextGraphs-9* (2014), 25.

[16] Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural models of factuality. *arXiv preprint arXiv:1804.02472* (2018).

[17] Roser Saurí. 2008. *A factuality profiler for eventualities in text*. ProQuest.

[18] Roser Saurí and James Pustejovsky. 2009. FactBank: a corpus annotated with event factuality. *Language resources and evaluation* 43, 3 (2009), 227.

[19] Roser Saurí and James Pustejovsky. 2012. Are you sure that this happened? assessing the factuality degree of events in text. *Computational Linguistics* 38, 2 (2012), 261–299.

[20] Sandeep Soni, Tanushree Mitra, Eric Gilbert, and Jacob Eisenstein. 2014. Modeling factuality judgments in social media text. *ACL* (2014).

[21] Hong Yu and Vasileios H. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *EMNLP*.