Sentiment Analisis with LSTM networks

Mildred Adolfina Nava Andaya
Faculty of Information, Media and Electrical Engineering
Technische Hochschule Köln
Cologne, Germany
mildred_adolfina.nava_andaya@smail.th-koeln.de

Yohanees Hutagalung
Faculty of Information, Media and Electrical Engineering
Technische Hochschule Köln
Cologne, Germany
yohanees.hutagalung@smail.th-koeln.de

Julio Cesar Ibañez Davila
Faculty of Information, Media and Electrical Engineering
Technische Hochschule Köln
Cologne, Germany
julio_cesar.ibanez_davila@smail.th-koeln.de

***Abstract* —Nowadays, millions of people use social media and webs services to comment on politicians and products. Sentiment analysis techniques classify data automatically based on polarity labels with the aim to improve products or predict the winner of an election. This paper proposes different models to detect sentiment polarity based in Long-Short Term Memory networks (LSTMs) which can deal with long term dependencies by introducing memory in a network model for prediction and visualization. The results show that the proposed models achieve accuracies ranging from 85% up to 92% being bidirectional LSTMs (BiLSTMs) the ones that achieved higher accuracies and showed reliability for further predictions. To improve this performance, it is proposed to train with more data. Complex sentences which express sarcasm were removed and are proposed for further investigation.**

Keywords—Twitter, Amazon, Sentiment, Classification, LSTMs, BiLSTMs.

I. Introduction

Social media such as Twitter and websites like Amazon are widely used by people to express their opinion on different topics, examples of these are product reviews and opinions about politics. Some companies benefit from this data to improve their earnings by sending product recommendations to users related to the product they reviewed [1]. Political Tweets are used by data analytics companies to keep track on regions where candidates are favourable and work towards regions where they are not in order to improve their chances to win an election [2], however is also possible to detect whether an election winner is legitimate or not based on the number of positive Tweets towards him.

Earlier deep learning features use Lexicon based techniques for supervised, unsupervised, and semi supervised approaches. Recent Deep Learning Architectures rely on powerful machine learning algorithms carefully designed for various domains from Text and Speech recognition, Natural Language Processing (NLP), to Computer vision, and even more [3]. Recurrent Neural Networks (RNN´s) along with Long Short-Term Memory (LSTM) classify Amazon and Tweeter data using keras and word embeddings, it has been shown that LSTMs are powerful RNN´s for sentiment analysis, LSTM units called memory cells learn long term dependencies from a sequence of words previously pre-processed.

Good results have been achieved predicting sentiments on consumer reviews however, political opinions pose a great challenge for LSTMs since validation accuracies of only 60% has been reached [4]. In this work 2 data sources will be used to train different LSTM networks and generate predictions to classify data as positive or negative; two data sources were taken for this research: consumer reviews from Amazon and political opinions from Tweeter. Different approaches will be implemented to generate good predictions. The proposed models will be train separately to predict sentiments only on product reviews or political tweets, after that both data sources will be mixed to train different LSTM architectures so that good results with both kind of data sources is achieved. Different LSTMs like stacked or bidirectional will be trained.

This paper comprises VII sections: Section I shows an introduction. Section II covers the literature related to the proposed work. Section III gives a short overview what is sentiment analysis with machine learning. Section IV describes the data to be used and all pre-processing techniques required to feed the LSTM models. Section V will explain in detail the methods and architectures implemented in this research. Section VI discusses about evaluation of experimental data whereas, Section VII concludes the overall work.

II. Related work

Many researchers have tried to implement sentiment analysis models for politic purposes. One of them has evaluated well-formed exploratory sentiment-based analysis of Twitter data that was gathered before and after the US Election Day on 2016 [5].

Twitter sentiment analysis is also being used in a vast array of areas related with governance and public trust, which range from predicting resentment against government policies to predicting general election results [6].

Various models that try to understand the user behaviour and “retweeting” on Twitter have been developed [7]. Other studies have looked at how information is diffused on social networks and what role sentiment plays on this. It is well-known and should be recognized that sentiment does play an important role in information diffusion on social networks [8].

Data pre-processing methods like Word tokenization, normalizing and stemming in conjunction with supervised machine learning models have been proposed [9], they achieved precisions of 92% classifying Arabic political tweets.

Reference [10] proposed sentiment analysis of US airline service using word embedding models (Doc2Vec) of six airlines with seven different features classifiers of KNN (KNearest Neighbour), AdaBoost, Support Vector Machine

(SVM), Decision Tree, Random Forest, Logistic regression, and Naïve Bayes classifier algorithm.

Finally, [11] proposed sentiment classification for short term from social media posts with word embedding models using LSTM approaches which show better improvement than Extreme Learning Machines (ELM) and Naïve Bayes classifier.

III. SENTIMENT ANALYSIS

Sentiment analysis is the contextual mining of text that identifies and extracts subjective information in the source material.

However, analysis of social media streams is usually restricted to just basic sentiment analysis and count-based metrics.

Sentiment analysis is an automated process of analyzing text data and classifying opinions as negative, positive, or neutral. Sentiment analysis is the most common text classification tool. It can be subtracted some attributes of expressions like polarity, subject or opinion holder and can be applied at different levels: document-level, sentence-level, and subsentence-level.

Fine-Grained sentiment analysis in "figure [1]" is implemented in this paper and categorizes opinions with polarity labels: Neutral, positive, negative. Emotion detection aims at detecting emotions such as happiness, frustration, anger, sadness, etc. Aspect-based sentiment analysis identifies aspects or features in opinions. Intent analysis basically detects what people want to do with text rather than what people say with that text.

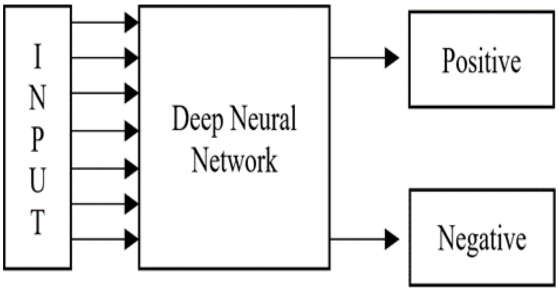

Fig. 1. Fine-Graded sentiment analysis

IV. ABOUT THE DATA SET AND TEXT PREPROCESSING

Two different data sets are taken to compare the accuracy of LSTM networks when predicting sentiments. The first data set consist of 400000 amazon reviews in figure "2" labelled as positive with a "2" and negative with a "1". As this amount of data can be overwhelming for normal computers it was split in 50% to be subsequently pre-processed and tokenized.

2	Great CD	My lovely Pat has one of the GREAT voices of her generation. I have listened to this CD for Y
2	One of the be	Despite the fact that I have only played a small portion of the game, the music I heard (plus
1	Batteries die	I bought this charger in Jul 2003 and it worked OK for a while. The design is nice and conven
2	works fine, b	Check out Maha Energy's website. Their Powerex MH-C204F charger works in 100 minutes f
2	Great for the	Reviewed quite a bit of the combo players and was hesitant due to unfavorable reviews and
1	DVD Player c	I also began having the incorrect disc problems that I've read about on here. The VCR still w
1	Incorrect Dis	I love the style of this, but after a couple years, the DVD is giving me problems. It doesn't ev
1	DVD menu se	I cannot scroll through a DVD menu that is set up vertically. The triangle keys will only select
2	Unique Weir	Exotic tales of the Orient from the 1930's. "Dr Shen Fu", a Weird Tales magazine reprint, is a
1	Not an "ultim	Firstly,I enjoyed the format and tone of the book (how the author addressed the reader). Ho

Fig. 2. Amazon Tweets.

The second data set is a collection of Tweets made by the US republican party taken from "Kaggle", another portion of Tweets taken from the same web site consisting of Trump insults. The tweets from the Republican party contain not only political opinions but also some general commentaries which are not useful, therefore only 800 samples were extracted and labelled with a "1" if they are positive and negative tweets were labelled with a "0". 500 Trump insults were taken and labelled in the same manner, finally both labelled tweets were joined summing up a total of 1300 tweets.

A. *Data preprocessing*

Before the LSTM can be fed, an important step must be taken, first the data must be cleaned. Computers cannot read text the way humans do. So, is important to reformat reviews in the dataset to make them easier for the system to understand. This means a special function was written to remove capitalization, punctuation, and articles (a, an, and the). This process is illustrated in figure "3".

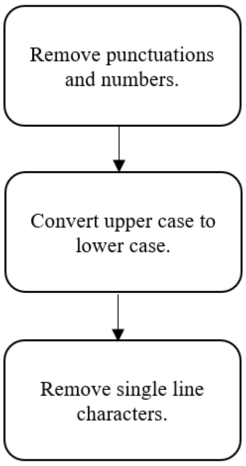

Fig. 3. Data pre-processing diagram

B. *Word vectorization*

The tokenizer is used to vectorize a text corpus (the text data being worked on), by turning the text into either a sequence of integers (each integer being the index of a token in a dictionary) or into a vector where the coefficient for each token could be binary or based on word count.

Three important functions will be used to tokenize data:

a) fit_on_texts(): Creates the vocabulary index based on frequencies. Every word gets a unique integer value.

b) texts_to_sequences(): takes each word in the text and replaces it with its corresponding integer value from the word_index dictionary.

c) Pad_sequences: Appends zero values to the tokenized list of words so that of lists have the same length.

An example of the output produce after this treatment is shown in "figure 4".

```
[[   597    50     3 ...    45   208    11]
 [41827 22031   206 ...   166    54     0]
 [    88   638    46 ...     0     0     0]
 ...
 [   550    48    21 ...   279   201    88]
 [    31   415  1629 ...     0     0     0]
 [    28   456   123 ...    48     0     0]]
```

Fig. 4. Vectorized text sequences

V. LSTM´S AND RESEARCH METHODOLOGY

A. Long short-term memory (LSTM) network

Learning to store information over extended time intervals by recurrent backpropagation takes a very long time, mostly because of insufficient, decaying error backflow [12]. An LSTM is capable of learning long-term dependencies. It maintains a relatively constant error that allows recurrent nets to continue to learn over multiple steps to link causes and effects remotely. This error is back-propagated through time and layers.

Figure "5" shows a typical 2 layered LSTM network where the blue squares are a memory cells, x_t, x_{t-1}, x_{t-2} are a sequence of inputs through time and ${c_t}^n, {c_{t-1}}^n, {c_{t-2}}^n$ are the outputs at the n layer.

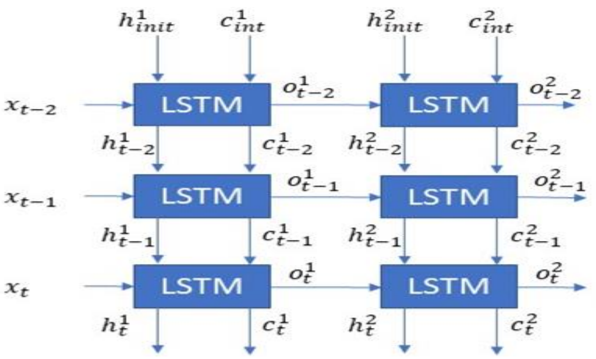

Figure 5. LSTM network with two layers.

A LSTM essentially acts as a memory cell in which information can be stored, written to, or read from. The cell makes decisions about what to store and when to allow read and write privileges via analogue gates. These gates are implemented with element-wise multiplication by Sigmoids.

The cells learn when to allow data to enter, leave, or be deleted through the iterative process of making guesses, back propagating error, and adjusting weights via a gradient descent. It provides a solution to the "vanishing gradient" problem by creating a connection between the forget gate activations and the gradients' computation. This connection creates a path for information flow through the forget gate.

a) Forget gate: The output of the forget gate tells the cell state which information to forget by multiplying a position in the matrix by zero. If the output of the forget gate is 1, the information is kept in the cell state. The equation for this gate is shown in equation (1).

$$f_{gt} = Sigmoid(W_{fg} \cdot [n_{t-1}, a] + b_{fg}) \quad (1)$$

Where W_{fg} is the weight matrix for the forget gate, a represent the inputs n_{t-1} is the output of the previous cell and b_{fg} is the bias for this gate.

b) Input gate: This gate determines which information should enter the cell state. This is done in a two-step process: A sigmoid function to decide which values will be updated (2) and a tanh layer creates a vector of the new values (3).

$$\mathrm{i_{gt}} = \mathrm{Sigmoid}(\mathrm{W_{ig}} \cdot [\mathrm{n_{t-1}}, \mathrm{a}] + \mathrm{b_{ig}}) \quad (2)$$
$$\mathrm{S_{t1}} = \tanh(\mathrm{W_c} \cdot [\mathrm{n_{t-1}}, \mathrm{a}] + \mathrm{b_c}) \quad (3)$$

Where $\mathrm{W_{ig}}, \mathrm{W_c}$ are the weight matrixes for the input gate and the cell state respectively and $b_{ig}, \mathrm{b_c}$ are the biases for the input gate and the cell state.

c) Update the state: the results of these two steps are combined to create an update to the cell state (4).

$$\mathrm{S_{t2}} = (\mathrm{f_{gt}} \cdot \mathrm{S_{t-1}}) + (\mathrm{i_{gt}} \cdot \mathrm{S_{t1}}) \quad (4)$$

Where $\mathrm{S_{t-1}}$ is the previous state.

d) Output gate: This gate takes into account all the possible values and utilizes a Sigmoid activation function to decide which information should go to the next hidden state (5).

$$O_{gt} = Sigmoid(W_{og} \cdot [n_{t-1}, a] + b_{og}) \quad (5)$$

O_{gt} represents the output gate, W_{og} and b_{og} are the weight matrix and the bias for this gate.

Finally, the cell output is obtained by multiplying the cell state and the result from the output gate (6):

$$\mathrm{n_t} = \mathrm{o_{gt}} * \mathrm{S_t} \quad (6)$$

The set of operations defined for the LSTM unit are illustrated in figure "6".

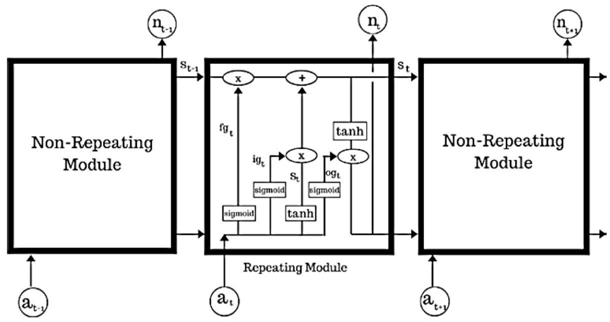

Fig. 6. Operations in a LSTM unit

B. Bidirectional LSTM (BiLSTM)

One shortcoming of conventional RNN´s is that they are only able to make use of previous context. Bidirectional LSTMs explore future context by processing data in both directions with two separate hidden layers, which are then fed forwards to the same output layer [13].

BiLSTMs show very good results as they can understand the context better by using the information from the future. Typically, two separate RNNs are used: one for forward direction and one for reverse direction. This results in a hidden state from each RNN, which are usually concatenated to form a single hidden state.

The final hidden state goes to a decoder, such as a fully connected network followed by SOFTMAX. Depending on the design of the neural network, the output from a BiLSTM can either be the complete sequence of hidden states or the state from the last time step. However, stacking many layers of BiLSTM creates the vanishing gradient problem.

Figure “7” illustrates a single layer BiLSTM network.

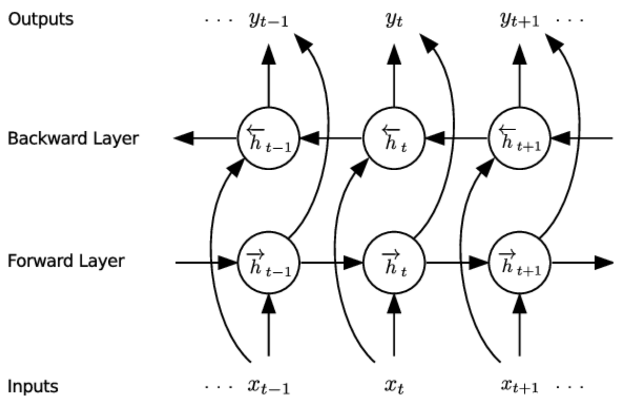

Fig. 7. Single layer LSTM network

C. Proposed architectures and metodology

To compare how well the data is classified by RNNs, different LSTM architectures will be trained separately with the amazon reviews and political tweets, then it will be compared with which kind of dataset the best results are obtained. Finally, both sources of data will be mixed to find if the network is able to learn from these and give an accurate positive or negative classification despite the source of data.

a) Embeding layer: Alows to translate high-dimensional vectors into a low dimensional space. It makes easier to implement machine learning on large sparce vectors representing words. Embedding captures some of the semantics of the input by placing semantically similar inputs close together in the embedding space.

b) LSTM layer: As the accuracy provided by the LSTM and BiLSTM networks is to be proved, it has been decided to experiment with a single LSTM layer and test if this is capable to provide high classification accuracy, then 2 stacked LSTM layers will be implemented to verify if there is an improvement in accuracy, in theory a BiLSTM have a better understanding of the context in a sentence, therefore a simple BiLSTM and a two layered BiLSTM are proposed to examine if they offer a better performance over simple LSTMs.

c) Dense layers: The final dense layer which will give the output to classify sentiments consists of a sigmoid activation function suited for binary classification, as sentiments are labeled with a “0” and “1” this is the ideal function for this research. It is proposed to stack a dense layer with a linear or SOFTMAX activation functions before the sigmoid layer. The linear activation function provides a better performance while the SOFTMAX function is a common function for classification problems and normalizes the output for each class between 0 an 1.

d) Dropout layers: To prevent overfiting a dropout layer will be placed at the input of the LSTM layer and also some LSTM cells will be dropped randomly, furtheremore it will be placed a dropot layer at every connecting dense layer. This approach force nodes within a layer to probabilistically take on more or less responsibility for the inputs.

D. Experimentation procedure

As this paper intents to compare how well different LSTM architectures classify political data and product reviews the steps to take for experimentation consist of the following:

Read already labeled data and subtract only 50% of them if they are the amazon reviews, political data and mixed data are ready for further processing. Then special functions will be coded to clean data, this means to remove special characters, single line words, and convert upper case to lower case letters.

After that, 70% of the cleaned data will split 70% for training and 30% for testing. A tokenizer in conjunction with an embedding layer will be applied to train the model. All models will be trained for 10 epochs to examine whether they need more training or less. After optimal results has been achieved for all models, the obtained loss and accuracy will be compared.

The proposed steps are shown in figure “[8]”.

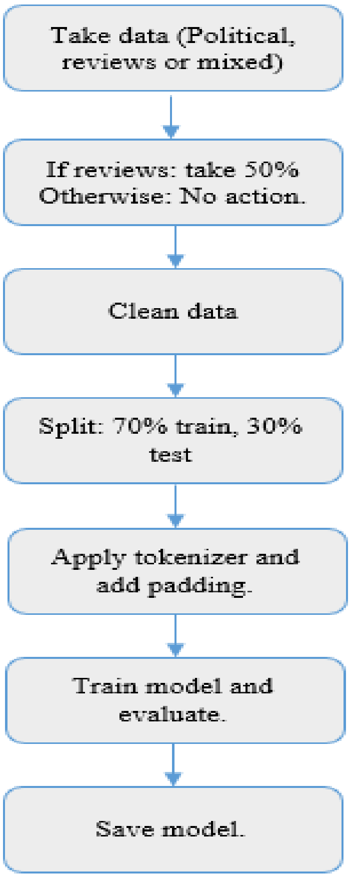

Fig. 8. Steps to take for experimentation.

VI. EVALUATION RESULTS

A. Results with amazon reviews

Different architectures will be trained to explore which one give the best results. The first candidate was a simple LSTM layer, the second model consists of 2 stacked layers, one third model is proposed with 4 stacked LSTM layers, finally a simple BiLSTM and a stacked BiLSTM network with 2 layers will be trained. Performance and validation accuracy of all models were compared and analyzed. Characteristics of all proposed architectures are resumed in table "1".

Model	**Layers**	**Dense layer 1**	**Dense layer 2**
1	1 LSTM	Linear	Sigmoid
2	2 Stacked LSTM´s	Linear	Sigmoid
3	4 stacked LSTM´s	Linear	Sigmoid
4	1 BiLSTM	Linear	Sigmoid
5	2 stacked BiLSTM´s	Linear	Sigmoid

Table 1. Different approaches to classify amazon reviews.

All models defined above were designed with 100 cells, a dropout of 0.5 at the input, a recurrent dropout of 0.5 and a dropout layer between each dense layer.

After all models where trained it was concluded that the 2 stacked layers BiLSTM network gave the best validation accuracy and the lower validation loss. The disadvantage this model presents is a slow training time due to the 8,954,817 trainable parameters, however this model achieved its maximum validation accuracy at the 3rd epoch. It can be concluded that this model provides a better understanding of the sentence context.

The single LSTM layer achieved the lower validation accuracy; however, it maintained an acceptable value of 0.8745. Model 2 reach better accuracy than model 1 and was also fast to train. Models 3 and 4 from the table, achieved similar results however model 4 had a better performance since it provided a fastest training time. Model 3 had the longer training time thus the worst performance.

Table "2" summarizes the results achieved by all models and provides the number of epochs required to reach their highest validation accuracy.

Model	**Validation accuracy**	**Epoch number**
1	0.8745	5
2	0.9088	4
3	0.9091	3
4	0.9098	4
5	0.9121	3

Table 2. Results obtained from classifying amazon reviews.

From table "2" can be inferred that LSTMs give very good results classifying reviews by sentiments. A simple BiLSTM layer offers the best trade-off between accuracy and performance and adding another layer to the model results in better accuracy, however stacking more and more LSTM layers lead to minor improvements while affecting performance. When accuracy is all it matters it can be said that stacking 2 BiLSTM layers is the superior choose. Model 5 is illustrated in figure "9".

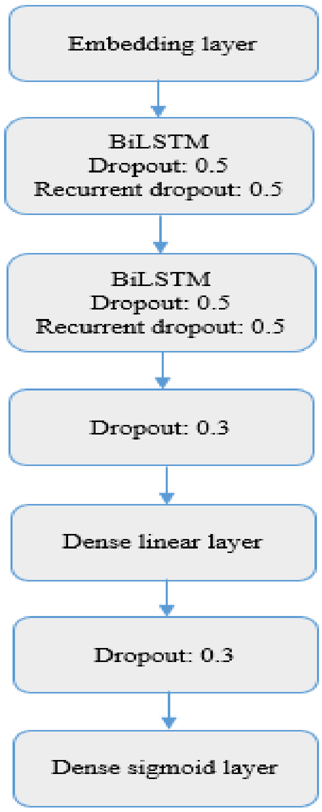

Fig. 9. Proposed BiLSTM model.

B. *Results with political tweets*

As previously mentioned, 1300 tweets where manually labeled and those which express sarcasm where removed, different approaches were proposed to get the highest possible validation accuracy, the decision was based on the results obtained with the amazon reviews, the model proposed are the following: 2-stacked LSTM layers, 1 BiLSTM layer and 2-stacked BiLSTM layers. As the amount of data was significantly smaller than the previous experiment it was decided to work only with 20 LSTM cells to prevent overfitting, the same dropout parameters were chosen. As with the previous experiment all models possess 2 dense layers, however the linear activation function from the penultimate dense layer was replaced by a SOFTMAX function.

The models proposed are displayed in table “3”.

Model	Layers	Dense layer 1	Dense layer 2
1	2 Stacked LSTM´s	SOFTMAX	Sigmoid
2	1 BiLSTM	SOFTMAX	Sigmoid
3	2 stacked BiLSTM´s	SOFTMAX	Sigmoid

Table 3. Proposed models to classify political tweets.

After training, the model with the simple Bidirectional layer achieved the highest validation accuracy, the lower validation loss and had the best performance. Followed by this, the model with 2-stacked LSTM layers outperformed model 3 in validation accuracy, validation loss and with fastest training time, this result is due to the amount of data employed to train the models, with only a small amount of data the approach with 2-stacked BiLSTM suffers from the vanishing gradient problem.

Table “4” summarizes the obtained results.

Model	Highest val_acc	Lowest val_loss	Epoch at lowest val_loss
1	0.9235	0.2737	93
2	0.9235	0.2698	100
3	0.9012	0.3061	100

Table 4. Results obtained from classifying amazon reviews.

The labels “val_acc” and “val_loss” corresponds to validation accuracy and validation loss, respectively. The plots which show graphically the accuracy and loss obtained during training and testing in the y-axis against epoch number in the x-axis for model 2 are shown in figures “10” and “11”.

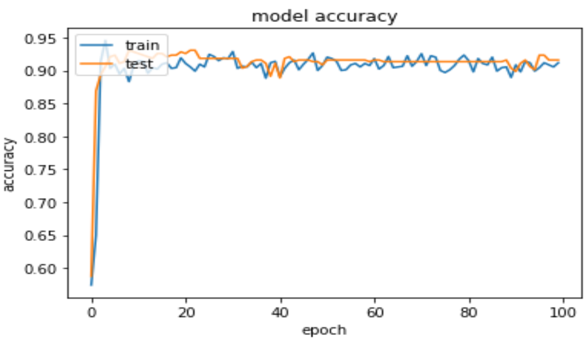

Fig. 10. Accuracy obtained in training and testing vs epoch number.

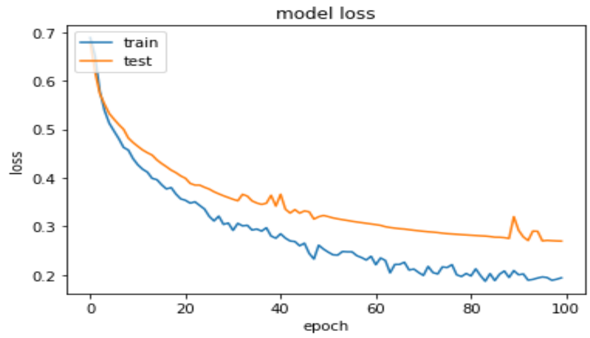

Fig. 11. Loss obtained in training and testing vs epoch number.

From the plots it can be concluded that the SOFTMAX function after the last dense layer can be beneficial with small amounts of data, it prevents overfitting and provides smoothness in accuracy and loss. Model 2 is illustrated in figure “12”.

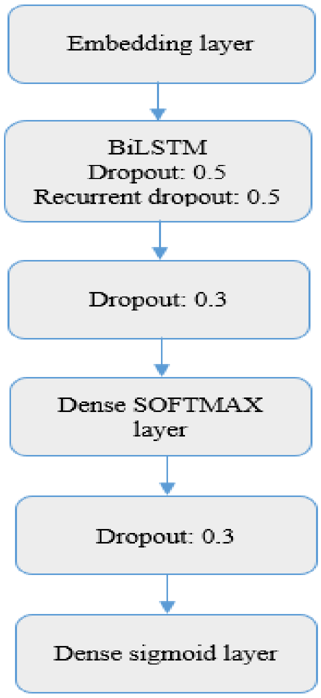

Fig. 12. Model proposed with 1 simple BiLSTM layer.

C. *Results with mixed data*

As the previous 3 models have reached high accuracy values, they will be used to classify product reviews mixed with political tweets so that they can learn to generalize positive and negative sentiments in any input text despite where it was collected. The number of input samples is 11300 for this reason a maximum 32 LSTM cells were chosen for all models; however, this number varies for those who possess stacked layers, table “5” summarizes how models were modified.

Model	Layers	Cells per layer	Dense layer 1	Dense layer 2
1	2 Stacked LSTM´s	32 first 16 second	Linear	Sigmoid
2	1 BiLSTM	32 one layer	Linear	Sigmoid
3	2 stacked BiLSTM´s	32 first 32 second	Linear	Sigmoid

Table 5. Models proposed for mixed data classification.

After the 1st epoch all models started to overfitting, different hyperparameters like dense layers, number of cells, dropout layers and batch sizes were tuned however they did not provide improvements in accuracy, this means only 1 epoch was enough for the models to learn the most important things from the data set. This time the SOFTMAX activation function did not help to give better improvements.

A maximum validation accuracy of 0.8607 was achieved by model 3, however all models reached similar results and showed a good performance. Table "6" summarizes the obtained outcomes.

Model	**val_acc**	**val_loss**	**Epoch at lowest val_loss**
1	0.8542	0.3592	1
2	0.8548	0.3267	1
3	0.8674	0.3240	1

Table 6. Results obtained from classifying mixed data.

From these results can be concluded that the data was just "to easy" to learn therefore the model can loss its accuracy classifying new data since it can face harder examples or many unknown words. Improvements to this can be done just by getting better or more data.

Model 3 is illustrated in figure "13".

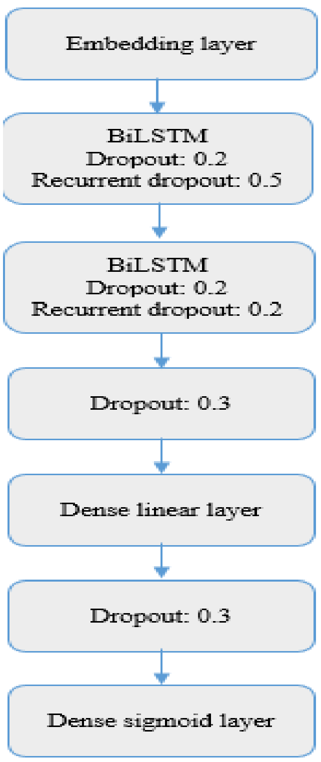

Fig. 13. Model proposed with 2 stacked BiLSTM layers.

D. *Predictions on new data*

The best models of the last three experiments were chosen to predict new experimental data with the aim to compare their accuracy with unknown sentences. These 3 models are summarized in table "7".

Model	**Layers**	**Cells per layer**	**Dense layer 1**	**Dense layer 2**
1	2 Stacked BiLSTM´s	100 first 100 second	Linear	Sigmoid
2	1 BiLSTM	20	SOFTMAX	Sigmoid
3	2 stacked BiLSTM´s	32 first 32 second	Linear	Sigmoid

Table 7. Models chosen to predict on new data.

From table "7" model 1 corresponds to me best model when classifying products reviews, model 2 reached higher accuracy with political tweets and model 3 was the best classifying mixed data.

A positive opinion on a video game was given to model 1 to evaluate its predictions on new data, only political data were inputted to model 2 to test its accuracy, finally model 3 was fed with a positive political opinion and a negative product review.

All models could predict with high accuracy their given input data, the results obtained are displayed in table "8".

Model	**Input data**	**Predicted value**	**Loss**
1	Positive product review	0.9512	0.0488
2	Negative political comment	0.0726	0.0726
	Positive political comment	0.8212	0.1788
3	Positive political comment	0.8429	0.1571
	Negative product review	0.3218	0.3218

Table 8. Results obtained after predictions.

Model 1 had the highest accuracy due to the volume of data with which it was trained, however the other two models performed very well, as model 2 was only trained with 1300 political tweets there are a lot of words not known, thus accuracy could be affected by this, the same can be said for model 3 since the data set used to fed model 1 is 20 times more, this problem can be tackled only by training with more data.

VII. CONCLUSIONS

This paper proposed different approaches to classify very relevant data on social networks and web sites like product reviews and political tweets. For large amounts of data, the model with 2 stacked BiLSTMs and two dense layers gave the best results however due to the number of trainable parameters this model was the second slowest to train. 4-stacked LSTM layers shows better accuracy than 2-stacked LSTM layers for large data sets nevertheless with small data sets it tends to overfit and the 2-layered LSTM approach outperforms it. Political data and product reviews can be classified with high accuracy if opinions which express sarcasm are removed; however, these are proposed to train a model specialized to classify sarcastic opinions as LSTMs showed a great potential to understand contexts. Making use of a SOFTMAX activation function in a dense layer was beneficial to prevent overfitting and provide smoothness on binary classification with small amounts of data employing LSTM´s. A linear activation in a dense layer before the last sigmoid layer provided very good results with large data sets. It was possible to classify political tweets and product reviews with the same model with acceptable accuracy thanks to key words like good, bad, worse, better, fraud or best, they are very general and are always used to describe an object or a subject. Finally, from all experiments realized the approaches which involved BiLSTMs achieved the highest accuracies. 1 simple BiLSTM layer is recommended for small data sets while 2-stacked BiLSTM layers are the best solution with high volumes of data.

REFERENCES

[1] Joseph A. Konstan, John Riedl, Barry Smyth, “ACM Conference on Recommender Systems,” Association for Computing Machinery, Minneapolis MN USA, October 2007.

[2] Harfoush, R ., “Yes we did! An inside look at how social media built the Obama brand,” New Riders, 2009.

[3] P. Vateekul and T. Koomsubha, “A study of sentiment analysis using deep learning techniques on Thai Twitter data,” 2016 13th Int. Jt. Conf.Comput. Sci. Softw. Eng. JCSSE 2016, 2016.

[4] Bakliwal, Akshat, Foster, Jennifer , van der Puil, Jennifer, O'Brien, Ron, Tounsi, Lamia and Hughes, Mark, “Sentiment analysis of political tweets: towards an accurate classifier,” NAACL Workshop on Language Analysis in Social Media, 13 June 2013, Atlanta, GA.

[5] Jaideep Vaidya, Soon Ae Chun, Ussama Yaqub, Vijayalakshmi Atluri, “Sentiment based Analysis of Tweets during the US Presidential Elections,” Conderence Paper, June 2017.

[6] Tumasjan, Andranik, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe, "Predicting Elections with Twitter: What 140

[7] Characters Reveal about Political Sentiment," ICWSM, pp.178-185, 2010.

[8] Broersma, Marcel, and Todd Graham, "Social media as beat: tweets as a news source during the 2010 British and Dutch elections," Journalism Practice 6, no. 3, pp. 403-419, 2012.

[9] Ferrara, Emilio, and Zeyao Yang, "Quantifying the effect of sentiment on information diffusion in social media," PeerJ Computer Science 1, 2015.

[10] Tarek Elghazaly, Amal Mahmoud, Hesham A. Hefny, “Political Sentiment Analysis Using Twitter Data,” Proceedings of the International Conference on Internet of things and Cloud Computing, pp. 1-5, March 2016.

[11] A. Rane and A. Kumar, “Sentiment Classification System of Twitter Data for US Airline Service Analysis,” Proc. - Int. Comput. Softw. Appl. Conf., vol. 1, pp. 769–773, 2018.

[12] J.-H. Wang, T.-W. Liu, X. Luo, and L. Wang, “An LSTM Approach to Short Text Sentiment Classification with Word Embeddings,” Conf. Comput. Linguist. Speech, 2018.

[13] Sepp Hochreiter, Jürgen Schmidhuber, “Long Short-Term Memory,” Vol. 9, No. 8, Neural Computation, Massachusetts Institute of Technology, March 13, 2006, pp. 1735-1780.

[14] A. Graves, N. Jaitly and A. Mohamed, "Hybrid speech recognition with Deep Bidirectional LSTM," 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, 2013, pp. 273-278.