A Deep Learning Approach to Aspect-Based Sentiment Prediction

Georgios Alexandridis, Konstantinos Michalakis, John Aliprantis, Pavlos Polydoras, Panagiotis Tsantilas, George Caridakis

A deep learning approach to aspect-based sentiment prediction

Georgios Alexandridis[1][0000−0002−3611−8292], Konstantinos Michalakis[1][0000−0002−5943−6613], John Aliprantis[1][0000−0001−5324−4103], Pavlos Polydoras[2], Panagiotis Tsantilas[2], and George Caridakis[1][0000−0001−9884−935X]

[1] Intelligent Interaction Research Group, Cultural Technology Department, University of the Aegean, University Hill, Mytilene, 81100, Lesvos, Greece
{gealexandri,kmichalak,jalip,gcari}@aegean.gr
https://ii.ct.aegean.gr/

[2] Palo Services, 9, Chavriou Street, 10562, Athens, Greece
{pp,pt}@paloservices.com
https://www.paloservices.com/

Abstract. Sentiment analysis is a vigorous research area, with many application domains. In this work, aspect-based sentiment prediction is examined as a component of a larger architecture that crawls, indexes and stores documents from a wide variety of online sources, including the most popular social networks. The textual part of the collected information is processed by a hybrid bi-directional long short-term memory architecture, coupled with convolutional layers along with an attention mechanism. The extracted textual features are then combined with other characteristics, such as the number of repetitions, the type and frequency of emoji ideograms in a fully-connected, feed-forward artificial neural network that performs the final prediction task. The obtained results, especially for the negative sentiment class, which is of particular importance in certain cases, are encouraging, underlying the robustness of the proposed approach.

Keywords: aspect-based sentiment analysis · bi-directional long short-term memory units · convolutional neural networks · attention mechanism · deep learning.

1 Introduction

Sentiment analysis or *opinion mining* has become a vigorous research area, especially in recent years, with the vast expansion of the *world-wide web* and the proliferation of *online social networks* (OSNs), like *Facebook*, *Twitter* and *Instagram*. Indeed, people discuss, voice opinions, share digital content and generally engage in activities, in a large public space. This reality has caught the attention of businesses and organizations, whose objective is to study and analyze public opinion with respect to the products and services they offer. Ideally, the aforementioned parties need not conduct surveys or opinion polls any more, as there is an abundance of relevant information available online.

However, locating and extracting user opinion from online sources (social media sites, blog posts, forums, etc) is a rather cumbersome task. Apart from the huge volume of information that needs to be processed, one has to be familiarized with the specifics of each service (e.g. API calls) and with the sentiment annotation processes. Therefore, it is not uncommon for companies to resort to specialized analysts that offer content services for consumers and brands.

From the business analyst perspective, sentiment analysis is a multi-faceted task. In [8], three distinct levels of analysis are identified; (i) document, (ii) sentence, (iii) entity and aspect. At the first level, a single sentiment is assigned on the whole document (e.g. positive or negative). This is practical for sources like news agencies, that usually discuss only one entity. At the second level of analysis, sentiment is extracted on a per sentence basis, having application on documents discussing more than one entities or on micro-blogging platforms like Twitter, where documents commonly consist of a few sentences.

The third level of analysis is the most demanding task, as instead of examining language constructs, the emphasis is placed on the entity or the aspect level. For example, a tweet stating *"Company X offers a great service, Thank God I switched over from Company Y"* can be classified as positive, w.r.t. Company's X service, negative w.r.t to Company Y and neutral, w.r.t. other similar companies. Therefore, the same text excerpt may have different interpretations. Additionally, subjective criteria may arise when deciding upon opinion or sentiment; for instance, a business may consider the reproduction of one of its press releases by a news agency a positive event, while another may view this event as neutral.

In this work, sentiment prediction is modelled as a supervised classification problem and is addressed using a deep learning architecture, based on *Bidirectional Long Short-Term Memory* (BiLSTM) units [5], combined with convolutional attention layers [10]. More specifically, Section 2 discusses related work and Section 3 presents the overall system architecture. Section 4 describes the data collected by the system, while Section 5 presents the feature extraction procedure, the implemented model and the obtained results. Finally, the work concludes in Section 6.

2 Related Work

Even though the research areas of sentiment analysis and opinion mining firstly appeared in 2003 [2], a multitude of works have been published on the subject ever since [8], based on various methodologies [11]. Nevertheless, in recent years, most state-of-the-art approaches are related to deep learning techniques. For example, the key element of the proposed system in [3] (studying the domain adaptation problem for sentiment classification) is a *stacked denoising autoencoder* that performs unsupervised feature extraction using both labeled and unlabeled samples. In [17], a neural network consisting of convolutional and LSTM layers is being presented, that learns document representations by considering sentence relationships. Other works combine the use of LSTMs with

attention mechanisms; for instance, in [20], the document-level aspect-sentiment rating prediction task is formulated as a comprehension problem that is being addressed by a hierarchical interactive attention-based model.

LSTMs have also been used in aspect level sentiment classification. In [16], the *target-dependent* and *target-connection* extensions to LSTM are proposed. The target is considered as another input dimension and is subsequently concatenated with the other features. A similar approach is followed in the current work; however, in the proposed methodology bi-directional LSTMs are employed instead. Bi-directional LSTMs with word embeddings at their input are used in [13] for aspect level sentiment classification, without an attention mechanism though.

An attention-based LSTM methodology for aspect-based sentiment analysis is described in [18], where the attention mechanism has been found to be effective in enforcing the model to focus on the important parts of each sentence, with respect to a specific aspect. Finally, in [19], two attention-based bidirectional LSTMs are proposed; however, unlike our approach, no other input features are considered apart from text.

3 System Architecture

The overall system architecture is depicted in Figure 1. It begins with *data crawling*, the process of systematically accessing a disparate set of online sources, gathering data that satisfy certain filtering criteria and forwarding them into a data repository. The set of data sources being accessed includes traditional web sources such as news sites, blogs, forums, as well as OSNs. There is a constantly updating registry of specific access points per medium, limiting search space only to relevant sources.

Each data source is correlated with potentially multiple data formats, depending on the information granularity. For instance, YouTube exposes hierarchical information, starting from a channel, drilling into metrics (followers and likes), its videos and finally video comments and reactions. The crawling process of this hierarchy needs to be addressed by the corresponding crawler, both in terms of data navigation and crawling policy. The latter is strongly related to data "freshness" (i.e. breaking news need to be crawled as fast as possible), as well as importance from a business perspective (e.g. more popular Instagram accounts need to be crawled more often). The crawling process, directed by these factors, pushes the accessed data sets to the ingestion services, in a streaming manner and re-iterates.

The subsequent step is data cleansing & homogenization. During this process, data are being stripped off inconsistencies attributed to major errors (e.g. missing article date), garbage information injection (e.g. ads in articles) and even erroneous semantics, such as out-of-scope or inappropriate content. Finally, data are stored in the *data lake* [12], a logical database which is the major hub of information exchange among the services. At its final version, the original raw data unit is upscaled into a mention, supplemented by derived information including

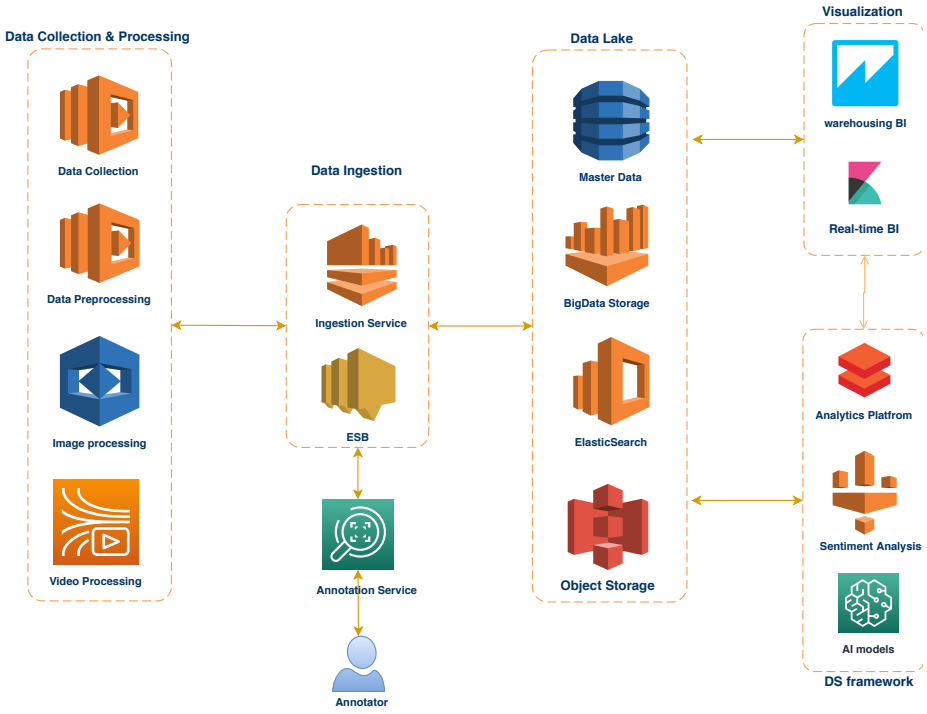

Fig. 1. Overall system architecture

named entities, image/pattern recognition measures, as well as sentiment indication. Mention semantics extend to a broader definition and usage context which includes a specific domain (e.g. telecommunications) and intended usage (e.g. competition analysis). Naturally, one document may correspond to multiple mentions, each associated with a different semantic context.

3.1 Annotation

Sentiment annotation [8] is the process of assigning specific sentiment values to a given mention. Currently, this process considers three values; namely *positive*, *negative* and *neutral*, but it can be generalized to a more extensive set. Orthogonal to sentiment assignment per se, however, is the knowledge base according to which a specific sentiment value can be extracted from a given mention. These rules can be arbitrarily chosen, based on specific criteria, driving sentiment analysis outcome accordingly.

In principle, sentiment annotation criteria are either *global* or *specific*, with the former referring to commonly agreed criteria, such as association of negative sentiment and lists of insulting words or phrases. The latter refer to specific rules, which may contradict the global ones and in those cases, they take precedence. The unit of focus is the aspect of the given mention being examined by the

rule. Aspects include the text of the mention and its metadata, which in turn consist of the respective data source information (e.g. news site and related category), entity type (e.g. Facebook post or comment), generation time (e.g. twitter comments posted after midnight), author, etc. In all cases, the aspect is well-defined prior to the annotating process and it is uniquely identified by a respective identifier (aspect id).

As stated above, one raw data record is associated with potentially multiple mentions, each corresponding to a different perspective. This results to possibly multiple sentiment values for the same record (remember the discussion about the different interpretations of the same tweet in Section 1), which are determined manually by a human annotator, studying the defined rules and applying them by assigning sentiment values to automatically selected samples. Sample selection follows the stratified random sampling methodology [14], with subgroups defined by the respective data sources (sites, blogs, social media, etc). Annotation is software-assisted and forms the respective data set (Section 4).

4 Data

Based on the procedure discussed above, 343,956 Greek language documents have been crawled and annotated over a period spanning nearly 2 years (September 2017 to June 2019). Table 1 outlines the distribution of their sources. As it is evident, the various sources are not evenly represented in this dataset because of the compliance of the crawling procedure to data protection regulations, as discussed in Section 3. For this reason, only public and business accounts are being processed and since Twitter is the most popular OSN in which information is disseminated predominately publicly, it is over-represented. The same reasoning is applied to news sources as well, as the vast majority of Greek news agencies and outlets are being monitored and indexed on a daily basis.

Table 1. Source medium distribution

Source	**Entries**	**Percentage**
Tweets	160,905	46.78%
News articles	87,750	25.51%
Facebook posts	39,591	11.51%
Facebook comments	25,250	7.34%
Blog posts	14,975	4.35%
Instagram	10,784	3.14%
Other	4,701	1,37%
Total	343,956	100.00%

Table 2 displays the frequency of appearance of specific domains within the crawled data. More than half of the collected information is about telecommunication businesses (mobile phone operators, Internet service providers, etc), followed by tobacco companies (about 20%). Another interesting observation is

the relatively large number of documents related to political parties and politicians, attributed to the fact that 2019 has been an election year in Greece. It should also be noted that the appearance of each specific domain is not evenly spread across all sources (that is, according to the distribution of Table 1). For example, information about the banking sector is predominately collected from news outlets ($\sim 70\%$, when news articles constitute a quarter of the dataset), while politics appear evenly on Twitter and on news articles ($\sim 50\%$ and $\sim 40\%$, respectively).

Table 2. Domain distribution

Domain	**Entries**	**Percentage**
Telecom	178,739	51.97%
Tobacco	72,822	21.17%
Banks	36,582	10.64%
Politics	30,677	8.92%
Retail	11,756	3.42%
Transport	8,079	2.35%
Misc	5,301	1.54%
Total	343,956	100.00%

Finally, Table 3 summarizes the distribution of the three categories of annotated sentiment (Section 3.1) over the whole dataset. In total, fiver persons participated in the annotation task, all of whom had received special training on annotation guidelines. Additionally, to further eliminate bias in the labels, well-defined annotation rules, such as cross-validation, irregular intervals and random data distribution (to all available annotators) across the dataset, have also been adopted.

As it is evident, it is highly imbalanced, since the *neutral* class is assigned to the overwhelming majority of the cases, while the other two (and especially the *positive* class) are clearly underrepresented. If sentiment distribution is further analyzed on a per source basis, the most negative content ($\sim 30\%$) appears on Twitter, a medium offering relatively anonymity (and thus, more "freedom") to its users. On the other hand, the least polarized opinions and at the same time the most "neutral" ones (more than 90%) appear on news articles. The latter are written by journalists who, most of the time, use a professional, unbiased language. Lastly, the most positive sentiment is expressed on Facebook comments ($\sim 12\%$), which is three times more than the average.

Table 3. Annotated sentiment distribution

Sentiment class	**Percentage**
positive	4.03%
neutral	78.53%
negative	17.44%

A similar analysis on a per domain basis is also interesting. By far, the most negative (and the least neutral or positive) feelings are expressed when politics are discussed, indicating that this is a highly polarized topic. The most neutral content, on the other hand, is again related to the banking sector, since most of the relevant content in the collected dataset originates from news articles, as it has been already argued. Finally, the transportation sector has received the most positive comments (around 10%).

4.1 Preprocessing

Prior to performing the sentiment prediction task, a number of data preprocessing steps are necessary. Initially, the textual part of each record is cleaned; that is, extra white space, non printable characters and other artifacts (e.g. HTML tags) are removed. Subsequently, the words that comprise the text are mapped to an *embedding space*, using *fastText* [1], a *natural language processing* methodology. In the end, the text of each document, is represented by the embeddings (vectors) of its words.

Among the non-printable characters that are extracted in the cleaning phase are *emojis* [9], a short of ideograms used in electronic communications to express feeling and emotions that are directly related to the sentimental state of the author of the document (e.g. smileys, sad or angry faces, etc). Since emojis do carry sentiment information, they are expected to positively contribute to the opinion mining task. A common methodology of including emojis in the prediction task would be to map them to a continuous vector space, usually consisting of two dimensions (sentiment score and neutrality) [9]. However, a different approach has been followed in this work; instead of using emoji embeddings, a vector designating the frequency of appearance of each emoji has been constructed for each record.

Another important characteristic that might be related to the sentiment value of a record is the number of its repetitions (retweets, shares, reposts, etc). The intuition behind this type of reasoning is that widely-spread content may carry significant emotional weight and therefore a correlation might exist between the number of times a text excerpt appears and its content. This characteristic follows a power law distribution in the dataset; the overwhelming majority of documents appear only once, while less than a $1,000$ records have been repeated more than 10 times. For this reason, in the model of Section 5.2, the logarithm of the number of repetitions is considered.

5 Experiments

The experiments that follow have been performed on the collected corpus presented above. In order to maintain temporal consistency, the dataset has been chronologically split into a training set (63.75% of the samples, earliest in time), a validation set (11.25% of the samples, subsequent in time) and a test set (25% of the samples, latest in time).

5.1 Feature Extraction

The predominant feature extraction activity involves the textual parts of each record in the collection. It is achieved by a stacked, two-layered BiLSTM network (Figure 2), which is considered to be among the state-of-the-art in capturing the spatial relationship between words and the order they appear in a text sequence [21]. The neural embeddings of the words are provided to the network in the order they appear in text, with a small amount of Gaussian noise ($\mu = 0, \sigma = 1$) added to them, as a regularization effect that reduces overfitting. After extensive experimentation, the optimal number of units for each layer have been determined to be 150, with *dropout* layers applied in-between them ($p = 0.3$) [15].

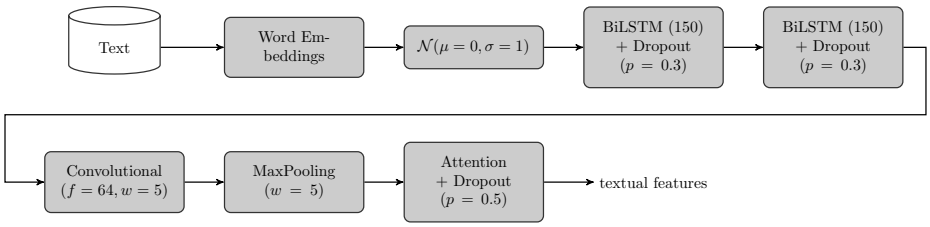

Fig. 2. Textual feature extraction procedure

After the BiLSTM layers, an one-dimensional convolutional layer follows, with 64 filters and a window size of 5. Again, both of the aforementioned hyper-parameters have been determined after experimentation. Subsequently, a max-pooling layer of an equal window size downsamples the output of the convolutional layer. The textual feature extraction is finalized with an attention layer, whose addition counterbalances the decline in performance when dealing with long sentences. Lastly, the feature extraction procedure concludes with a Dropout layer ($p = 0.5$).

The other three features to be considered do not require such an extensive feature extraction procedure. Aspect is incorporated through aspect id, an one-hot encoded variable (Section 3.1), while the presence of emojis is quantified as a frequency vector. Finally, the number of repetitions of each document is provided to the model via its logarithm (Section 4.1).

5.2 Model Selection

After experimenting with various techniques and architectures, the optimal model has been determined to be a fully-connected feed-forward artificial neural network, consisting of two hidden layers (Figure 3). The first hidden layer is comprised of 1024 neurons and the second of 128. Their activation function is the *rectified linear unit* [6] (in contrast to the output layer, where *softmax* activation is used instead [4]). Network training has been based on the Adam optimization

algorithm [7], with a learning rate of 10^{-3} and hyperparameters β_1, β_2 being fixed at 0.9 and 0.999, respectively. Finally, the *fastText* word embedding vectors used in the experiments have been pretrained on a corpus of more than $2,000,000$ words.

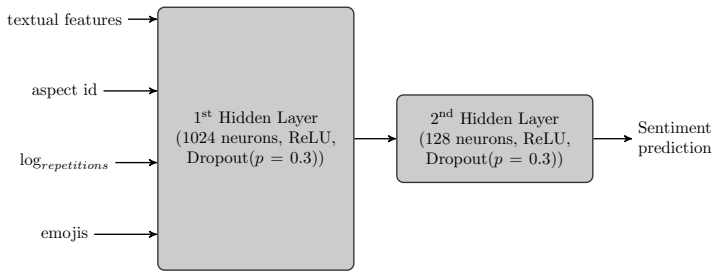

Fig. 3. Fully-connected model

5.3 Results

Figures 4-5 examine system performance regarding various aspects on *Precision, Recall* and their harmonic mean (*F1-score*); a set of popular information-retrieval metrics, widely used in sentiment analysis tasks [8]. The former is equal to the ratio of the correctly classified documents to a given class over the total number of classified documents to that class, while the latter is equal to the ratio of the correctly classified documents to a given class over the total number of documents that belong to that class. The values displayed in Figure 4 are averaged over all classes, while the results in Figure 5 are given for each class separately.

Figure 4 summarizes system performance with respect to the different inputs. When only the textual part of each record is considered, the efficiency of the proposed approach is limited, an indication that in the environment described in Section 3, text alone is not a sufficient indicator for the prediction task. When aspect-related information is considered, Recall increases by more than 6% (followed by a smaller boost in Precision), meaning that the system can better discriminate in-between the classes. The addition of the logarithm of the number of repetitions marginally affects Recall, but further enhances Precision, with the F1-score in this case being slightly better than the previous one. Finally, when all four inputs are provided (text, aspect, logarithm of the number of repetitions, emojis), system throughput is further enhanced, with all three metrics being above (70%), adding more than 8% to the overall system performance.

Figure 5 displays the per-class performance of the examined metrics. Even though class labels are highly imbalanced (Table 3), the system achieves very good results for the negative class, a characteristic that is of significant importance, as the main concern of many businesses is to be able to timely identify and respond to unpleasant content. On the other hand, the predictions on the

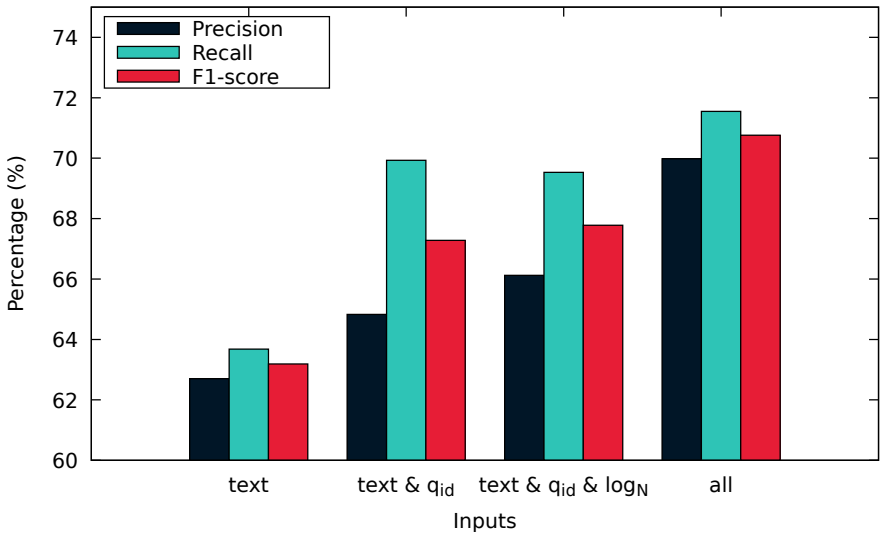

Fig. 4. System performance w.r.t different inputs

positive class are clearly below average and therefore more effort should be put in the direction of improving system efficacy for this particular case, as well.

6 Conclusions

In this work, a novel hybrid bi-directional LSTM/CNN feature extraction architecture has been presented, as part of a broader system that performs aspect-based sentiment analysis. The obtained results, on a corpus selected from Greek-language content from OSNs and other sources, are encouraging, especially on the negative class that is of particular interest to businesses. Nevertheless, the outlined architecture needs to be further fine-tuned and reasoned upon, as the system demonstrates sub-optimal performance in identifying positive sentiment.

The proposed architecture may be extended in a number of ways. An obvious direction would be to consider additional textual characteristics that convey aspect and sentiment-based information. For instance, feature extraction from hashtags, which are quite popular on OSNs, is expected to further aid the desired task. Additionally, the number of repetitions could be leveraged by examining frequency patterns in-between the users that are either mentioned on or just redistribute content.

Finally, the quality of the already extracted features may be farther enhanced. For example, the application of dimensionality reduction techniques, such as principal component analysis, on the emoji frequency matrix can help determine which of the available emojis have the greatest impact on the sentiment analysis task.

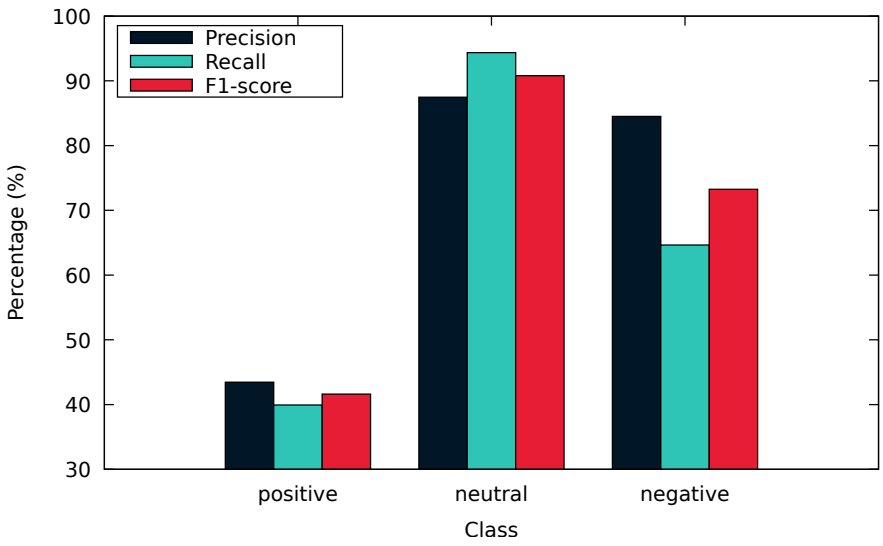

Fig. 5. System Performance w.r.t different classes

Acknowledgements

This research has been cofinanced by the European Regional Development Fund of the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH - CREATE - INNOVATE (project code: T1EDK-03470).

References

1. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)
2. Dave, K., Lawrence, S., Pennock, D.M.: Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In: Proceedings of the 12th International Conference on World Wide Web. pp. 519–528. WWW '03, ACM, New York, NY, USA (2003). https://doi.org/10.1145/775152.775226, http://doi.acm.org/10.1145/775152.775226
3. Glorot, X., Bordes, A., Bengio, Y.: Domain adaptation for large-scale sentiment classification: A deep learning approach. In: Proceedings of the 28th International Conference on International Conference on Machine Learning. pp. 513–520. ICML'11, Omnipress, USA (2011), http://dl.acm.org/citation.cfm?id=3104482.3104547
4. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016), http://www.deeplearningbook.org
5. Graves, A., Fernández, S., Schmidhuber, J.: Bidirectional lstm networks for improved phoneme classification and recognition. In: Duch, W., Kacprzyk, J., Oja,

E., Zadrożny, S. (eds.) Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005. pp. 799–804. Springer Berlin Heidelberg, Berlin, Heidelberg (2005)
6. Hahnloser, R.H., Sarpeshkar, R., Mahowald, M.A., Douglas, R.J., Seung, H.S.: Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. Nature **405**(6789), 947 (2000)
7. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6980
8. Liu, B.: Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers (2012)
9. Novak, P.K., Smailović, J., Sluban, B., Mozetič, I.: Sentiment of emojis. PloS one **10**(12), e0144296 (2015)
10. Raffel, C., Ellis, D.P.W.: Feed-forward networks with attention can solve some long-term memory problems. CoRR **abs/1512.08756** (2015), http://arxiv.org/abs/1512.08756
11. Ravi, K., Ravi, V.: A survey on opinion mining and sentiment analysis: Tasks, approaches and applications. Knowledge-Based Systems **89**, 14 – 46 (2015). https://doi.org/https://doi.org/10.1016/j.knosys.2015.06.015, http://www.sciencedirect.com/science/article/pii/S0950705115002336
12. Research, C.: Putting the data lake to work: A guide to best practices. Tech. rep., Teradata (2014)
13. Ruder, S., Ghaffari, P., Breslin, J.G.: A hierarchical model of reviews for aspect-based sentiment analysis. arXiv preprint arXiv:1609.02745 (2016)
14. Särndal, C.E., Swensson, B., Wretman, J.: Model assisted survey sampling. Springer Science & Business Media (2003)
15. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**, 1929–1958 (2014), http://jmlr.org/papers/v15/srivastava14a.html
16. Tang, D., Qin, B., Feng, X., Liu, T.: Effective lstms for target-dependent sentiment classification. arXiv preprint arXiv:1512.01100 (2015)
17. Tang, D., Qin, B., Liu, T.: Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings of the 2015 conference on empirical methods in natural language processing. pp. 1422–1432 (2015)
18. Wang, Y., Huang, M., Zhao, L., et al.: Attention-based lstm for aspect-level sentiment classification. In: Proceedings of the 2016 conference on empirical methods in natural language processing. pp. 606–615 (2016)
19. Yang, M., Tu, W., Wang, J., Xu, F., Chen, X.: Attention based lstm for target dependent sentiment classification (2017), https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14151
20. Yin, Y., Song, Y., Zhang, M.: Document-level multi-aspect sentiment classification as machine comprehension. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2044–2054 (2017)
21. Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., Xu, B.: Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. In: COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan. pp. 3485–3495 (2016), http://aclweb.org/anthology/C/C16/C16-1329.pdf