Question 1

MACHINE LEARNING ASSIGNMENT 3

Joshua Lake 12576930

The rise of social media in our day to day lives is creating connections and data like never seen before. It can be considered a data mining problem that needs to be solved in order to effectively make use of this plethora of data. We define data mining as the identification of patterns in data for knowledge discovery through the use of machine learning techniques. (Barber & Liu 2011) One such application of this can be considered the use of social network analysis in the form of collecting user messages, and whether it is possible to determine a user's sentiment, take this sentiment and either reinforce it or transform it. However, solving this problem gives rise to a variety of problems specific to the field of social media analysis (SMA).

Challenges to solving this problem

The first of these challenges comes down to the data itself, and the heterogenic nature of the population that is represented. (Faraset et al. 2015) These large social networks contain heterogeneous nodes and links, where there are multiple entities and relationships, adding a layer of complexity to an already large data visualisation. This makes reproducing and generalizing from statistical analysis reliant on how well the sample population is representative of the true population.

Secondly, the data collected is often noisy and includes grammatical errors, missing punctuation and slang. Along with this, the data includes ambiguity in terms of emotions being portrayed and requires the modelling of compositional sentiments through manipulation and identification of sentiment shifting words.

Finally, and perhaps the biggest issue in social media analysis, is privacy. The protection of data is of great concern for users of social networks, with a general preference to keep personal information to friends and family. (Faraset et al. 2015) However modern-day technology such as machine learning, has made it easy to invade privacy through the inference of personal attributes. The very basis of social media mean that social network analysis is reliant on how much people are willing to disclose to one another.

The Idea

A method proposed to make use of user messages for a polling organisation will combine aspect-opinion relation extraction with an SVM using an augmented dependency tree to perform sentiment analysis. This sentiment analysis will classify the users opinion on a given topic as either 1, neutral, or 0, and from this, the polling organisation can direct advertisements towards the classified users in order to influence their opinion.

The data will be collected in the form of text by querying various social media sites for keywords in relation to the topic being investigated. This can be as simple as inputting keyword queries or more complex such as using an indexing service to conduct real time indexing of text. An example of this is Spinn3r API, accessed through an open source Java application to mine through over 100,000 new pages of data per hour. This was effectively

used in text mining to reveal links between social media and real-world influenza-like symptoms (Corley 2009)

Following the collection of data, the data will undergo named entity recognition, in which it is organised into pre-defined categories. This is a natural language processing technique that is the first step in building the augmented dependency tree. This can be achieved using a toolkit such as the natural language toolkit (NLTK), which applies word tokenization and speech tagging to produce a list of tuples (A sequence of immutable objects). (Li 2018) This lays as a foundation to the pre-processing of the data, as such seen by Bjokerlund (2014), in which the API "openNLP" was used for entity tagging in the pre-processing stage of the building of a syntactic and semantic dependency parser, similar to the proposed model below.

As the data is now tagged, relationship instances can be generated by iterating over all pairs of entities occurring in the same sentence. This builds on work by (Culotta 2004) , where augmented dependency trees were created from each entity pair to create a training set of relations, and a tree kernel was used in an SVM to classify test instances. The dependency tree is built off the parse tree of each sentence. An example of an entity pair being represented as a dependency tree is seen below, where the entity pair is [movie, interesting]

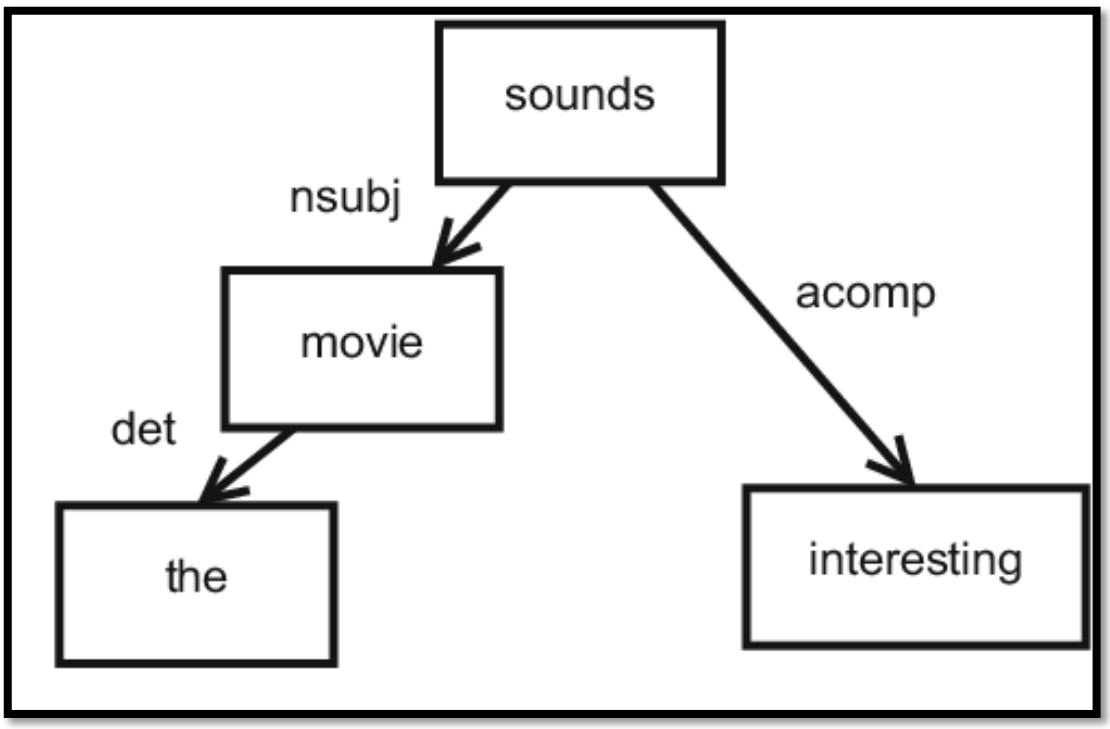

Figure 1: (A simple example of a dependency tree that has been tagged)

Each entity relation is then assigned a feature vector containing a variety of features that will assist in classification:

Word	The words in the entity relation
Word Type	Adjective, verb, noun etc
Entity Type	Person, place etc
Entity Level	Name, Number, Pronoun etc
Tag	NP, VP, ADJP, NNP etc
Polarity	1 or -1 or Net Polarity

Figure 2: (The feature vector)

One such feature, in which we draw upon work by Yasavur et al. (2014) , is the polarity of the entities. The sentiment polarity recognition is based on the notion that a sentence may contain positive or negative words, but the sentence as a whole may not be reflective of the entity. (Yasavur et al. 2014) uses a simple algorithm to determine polarity; if the polarity of a word is positive, it adds 1 and if it is negative it subtracts 1. The net polarity is taken to identify the polarity of the relation.

As suggested by (Zelenko 2003) a tree kernel function can then be defined in terms of a similarity function and a matching function. The kernel will compare two dependency trees in terms of similarity and matching, with the first tree being an aggregate of all training subtrees, and the second tree being the tree under investigation. The SVM will return a symmetrical similarity score between 1 and 0. The similarity function compares the nodes of the tree directly, whereas the matching function compares the feature vectors. We see this in Borele (2016), where semantic analysis was conducted using an SVM, with the classification algorithm firstly calculating the "centroid vector for every training class" before calculating the similarities between these centroids and a new document are used to assign a class.

If the return of the SVM is "1", then there is a positive relation between the aggregate tree and the test tree. If the return is "0", then there is a negative relation between the aggregate and the test tree. Based on this, the polling company can advertise accordingly based on the score; a score of 1 will see the user targeted with information to reinforce their opinion where as a score of 0 will see the user targeted with advertisements to change their opinion.

Alternatives

An alternative method proposed is the design of a neural network system that uses back propagation to classify social media text data. The data is pre-processed and tagged in a similar method as above, and the classification is into positive, neutral and negative clusters.

Back propagation takes the errors of the neural system to update the weights of the existing neural network with the aim to reduce errors in subsequent predictions. This can be seen in the below back propagation system design by Sharma & Mandoi (2018)

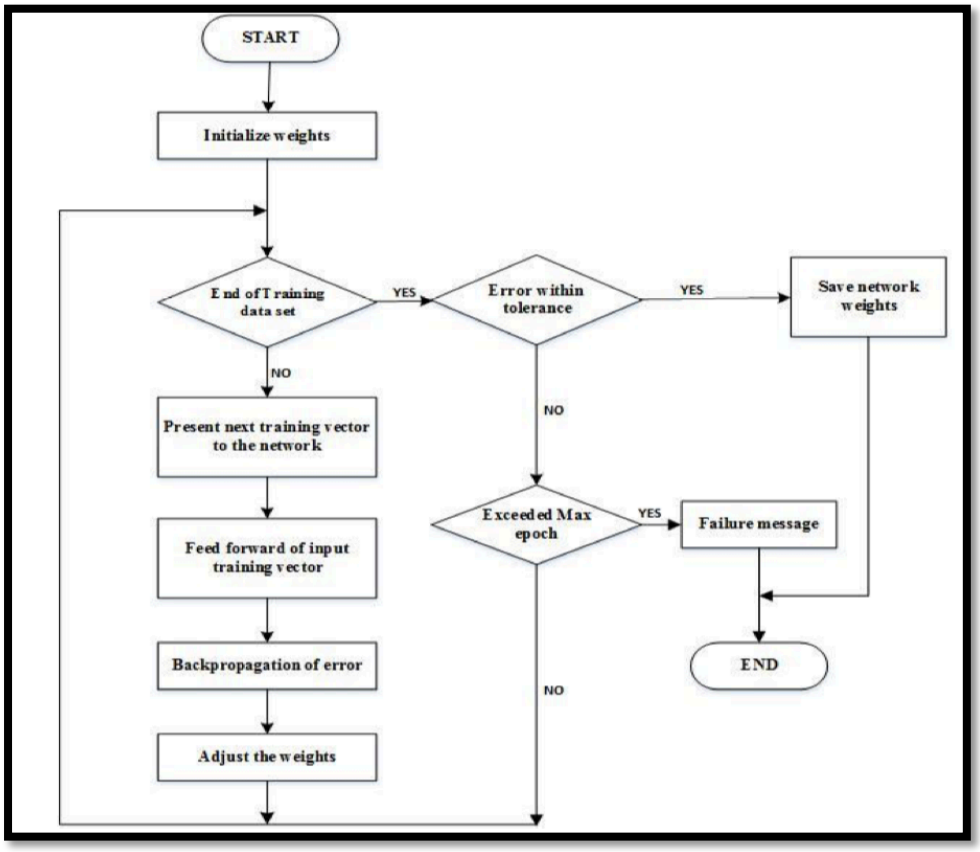

Figure 3: (Back propagation in an ANN)

A secondary alternative would be to use a multinominal Naïve Bayes classification algorithm to find the probabilities of classes assigned to text data by assessing the join probabilities of the words in the classes as suggested by Smetanin (2018). Knyszewski (2019) builds on this by proposing an algorithm that loops over sets of vocabulary to get a total word count (c), before computing the likelihood of each word.

The suggested SVM method however is superior to both the aforementioned methods as we have created a linear solution through the addition of the “polarity” feature in the vector of features. This allows for, in the 2-dimensional case, a hyperplane that separates the data points without any misclassification minimal overfitting.

Ethical Study

Study into this field however does raise the question of ethical and social consequences. As the study is based on public information, it can be seen as observational research. When collecting an individual's information, access must is deemed public if the information is identifiable but not private. Data cannot speak for itself, and we must consider the use of data in our model and where it is being sourced from. As online communications grow, social media devices and web applications are becoming ubiquitous, and the data available to all is becoming more readily available.

When users establish there "privacy settings" on their given social media platform, many of the details are often hard to find and are only located in fine print, without an emphasis on the so-called data sharing or collection of their information. As a result, there are many people who do not know that their information is in fact being collected. All social media platforms are required by law to have an option for the user to opt out of such data collection, however not all are aware of this.

Similarly, collecting individuals data for research purposes may be seen as beneficial and positive, however beyond the researcher, the data may fall into the wrong hands. If individuals private information makes its way into the wrong hands, there is no limit to the negative effects this could potentially have on the lives of the users. From bank crime to identity theft, it is becoming harder and harder to track where the data ends up.

Finally, using individuals information in order to tailor advertisements to the need of the polling organisation may be seen as unethical. Masters (2019) identifies 3 facets of ethically using social media research to tailor advertising:

1. Market researchers have an ethical obligation to conduct research that is objective, and the available data allows for the development of a balanced perspective.

2. Companies are required by law to not share information about customers or affiliates.

3. Companies must consider the users personal freedom when targeting advertisements

Furthermore, the above proposal will provide a fitting model for polling organisations to analyse messages collected from social media to predict how ones support for a particular topic can be changed. Through this, the social and ethical consequences have also been analysed to ensure the solution is both feasible and ethical.

References

A.Farasat, A.Nikolaev, S.N.Srihari & R.Hageman-Blair, 'Probabilistic Graphical Models in Modern Social Network Analysis', *Social Network Analysis and Mining,* Vol. 62. No.18, 2015

A.Culotta & J.Sorensen, 'Dependency Tree Kernels for Relation Extraction', *Proceedings of the 42nd Annual Meeting of the Association for Computation Linguistics,* July 2004

B.Agarwal & S.Porkia, 'Concept-Level Sentiment Analysis with Dependency-Based Semantic Parsing: A Novel Approach, *Cognitive Computation,* 2014

B.Bohnet, L.Hafdell & P.Nugues, 'A High Performance Syntactic and Semantic Dependency Parser', *Demonstration*, Vol.1, No.1, Beijing, August 2019

Borele & DA.Borikar, 'An Approach to Sentiment Analysis using Artificial Neural Network with Comparative Analysis of Different Techniques', *Journal of Computer Engineering,* Vol.18, Issue 2, No.5, April 2016

C.Corley, 'Text and Structural Data Mining of Influenza Mentions in Web and Social Media', *Int. J. Environ. Res. Public Health,* November 2009

D.Zelenko, C.Aone & A.Richardella, 'Kernel Methods for Relation Extraction', *Journal of Machine Learning Research, Vol.3,* 2003

G.Barbier & H.Liu, 'Data Mining in Social Media', *Social Network data Analytics,* Springer Science and Business Media 2011

Li.S, 2018, 'Named Entity Recognition with NLTK and SpaCy', *Towards Data Science: Named entity recognition*

A.Bjorkelund, N.Sharma & S.Mandloi, 'Sentiment Analysis of Social Media Text Data using Back Propagation in Artificial Neural Networks', I*nternational Journal for Research in Applied Science & Engineering Technology (IJRASET)* Vol.6, Issue.1, January 2018

S.Smetanin**,'** Sentiment Analysis of Tweets using Multinomial Naive Bayes', *Towards Data Science, September 2018, <* https://towardsdatascience.com/sentiment-analysis-of-tweets-using-multinomial-naive-bayes-1009ed24276b*>*

T.Masters,'Ethical Considerations of Marketing Research', *Small Businesses, March 2019,* https://smallbusiness.chron.com/ethical-considerations-marketing-research-43621.html

U.Yasavur, J.Travieso, C.Lisetti & N.Rishe,'Sentiment Analysis using Dependency Trees and Named-Entities', *Proceedings of the Twenty-Seventh International Florida Artificial Intelligence Research Society Conference,* 2014